DOCUMENT RESUME

ED 442 245 EC 307 883

AUTHOR	Tindal, Gerald; Fuchs, Lynn
TITLE	A Summary of Research on Test Changes: An Empirical Basis for Defining Accommodations.
INSTITUTION	Mid-South Regional Resource Center, Lexington, KY.
SPONS AGENCY	Special Education Programs (ED/OSERS), Washington, DC.
PUB DATE	2000-03-00
NOTE	125p.
CONTRACT	H326R980003
AVAILABLE FROM	Information Services, Mid-South Regional Resource Center, University of Kentucky, 126 Mineral Industries Bldg., Lexington, KY 40506-0051; Tel: 606-257-4921 (Voice); Tel: 606-257-2903 (TTY); Web site: http://www.ihdi.uky.edu/MSRRC (may be copied, document not copyrighted).
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Accountability; *Disabilities; Distractors (Tests); Educational History; Elementary Secondary Education; *Individualized Education Programs; *Performance Factors; *Response Style (Tests); Standardized Tests; Student Evaluation; Test Format; *Test Validity; Test Wiseness; *Testing; Timed Tests
IDENTIFIERS	*Testing Accommodations (Disabilities)

ABSTRACT

This document summarizes the research on test changes to provide an empirical basis for defining accommodations for students with disabilities. It begins by providing an historical overview of special education accountability. It describes how separate special education accountability systems have evolved and summarizes information on the participation of students with disabilities in general education accountability systems. The role of the Individualized Education Program as the main vehicle for expressing the need for test accommodations is emphasized. The paper then summarizes the research on test changes using a taxonomy from the National Center on Educational Outcomes. Testing accommodations are reviewed relating to timing and scheduling of testing, test settings, computer presentation of tests, examiner familiarity, multiple changes in presentation, dictation to a proctor or scribe, using an alternative response, marking responses in test booklets, working collaboratively with other students, using word processors, using calculators, reinforcement, and instruction on test-taking strategies. The last section of the document addresses issues of validity with primary considerations on using this research to implement sound testing practices and to make appropriate educational decisions. (Contains over 170 references.) (CR)

MID-SOUTH REGIONAL RESOURCE CENTER

IDEAs that Work
U.S. Office of Special Education Programs

A Summary of Research on Test Changes: An Empirical Basis for Defining Accommodations

By Gerald Tindal, Ph.D.
University of Oregon
And
Lynn Fuchs, Ph.D.
Vanderbilt University

July 1999
(Revised March, 2000)

Commissioned by the
Mid-South Regional Resource Center
Interdisciplinary Human Development Institute
University of Kentucky
126 Mineral Industries Building
Lexington, Kentucky 40506-0051
606-257-4921
Fax: 606-257-4353
TTY: 606-257-2903
www.ihdi.uky.edu/MSRRC

This document was developed pursuant to cooperative agreement #H326R980003 under CFDA 84.326R between the Mid-South Regional Resource Center, Interdisciplinary Human Development Institute, University of Kentucky and the Office of Special Education Programs, U.S. Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the U.S. Office of Special Education Programs and no endorsement by that office should be inferred.

TABLE OF CONTENTS

Abstract

This document summarizes the research on test changes to provide an empirical basis for defining accommodations. We analyze this research from three perspectives:

- Tests are changed in specific ways in the manner that they are given or taken.
- The change does not alter the construct of what is being measured.
- The changes are or can be referenced to individual need and differential benefit, not overall improvement.

In this review, a very wide sweep of the literature was made, using many key words to search both electronic databases and educational journals. Although the main focus was on test changes for students with disabilities, the literature was not confined to only studies done with this population. In fact, test accommodations can and should be studied in the context of validity, which implies both measurement and decision-making. Clearly such decision-making occurs in both general and special education. Using the latest amendment to the Individuals with Disabilities Educational Act 1997, we consider test changes as part of inclusion and progress in the general education curriculum. The first section addresses Individualized Education Programs (IEPs), using this component as the main vehicle for expressing the need for test accommodations. Then the research is summarized using a taxonomy from the National Center on Educational Outcomes (NCEO). The last section addresses issues of validity with primary consideration on using this research to implement sound testing practices and to make appropriate educational decisions.

Forward and Acknowledgments

This document was possible only through the careful and diligent work of many people who made significant contributions. First and foremost, Ken Olsen must be recognized, not only for his insight and vision in conceiving of this document as a summary reference for bridging research into practice but also for his very keen and helpful editing of several editions. Also, Brian Megert and Bill Raes must be recognized for their literature search and abstracting of the articles. As Master's degree students in a teacher preparation program, they went well beyond any requirements for licensure and entered the profession as scholars in training with their work on this document. Finally, Raina Megert needs to be given recognition for her editing and formatting of the document, especially for the many editions that, at the time, appeared to be the final and last version. Time after time.

As this document was being finalized, Patricia Almond, Oregon Department of Education, enlisted a study group from the Assessing Special Education Students (ASES) State Collaborative on Assessment and Student Standards (SCASS). She organized several individuals to serve as helpful critics and consumers, helping move this document to a form that would serve a very practical purpose. Following are the individuals who participated in this review:

Sue Bechard-Colorado Department of Education

Merrie Darrah- East Shore SERCC

Edna Duncan-Mississippi Department of Education

John Haigh-Maryland Department of Education

Cherie Randall Mercer-Kansas Department of Education

Nancy Maihoff-Delaware Department of Education

Tom Schoeck-New York Department of Education

Martha Thurlow-National Center on Educational Outcomes

ACCOUNTABILITY AND INDIVIDUAL EDUCATION PROGRAMS

Over the past decade, public demand for accountability for student outcomes has increased. With this increasing demand, reliance on large-scale student assessments has also increased. This demand has been codified in national education reform legislation, Goals 2000: Educate America Act, which requires national groups to oversee educational goals, standards, and assessments. Moreover, high-stakes consequences, such as grade promotion and awarding of high-school diplomas, increasingly are attached to statewide assessments.

Within special education, accountability has evolved separately from the frameworks employed within general education. Recently, however, questions have been raised about the tenability of accounting for special education outcomes outside the general education system.

In this introductory section, we provide an historical overview of special education accountability. We begin with a description of why and how a separate special education accountability system has evolved. Then, we summarize information on the participation of students with disabilities in general education accountability systems. Finally, we describe current initiatives to broaden public accountability frameworks so that they include all children, including those with disabilities.

Why and How Special Education Accountability Evolved As a Separate System

At least three reasons explain the evolution toward a separate special education accountability system. First, students with disabilities historically have been excluded from general education accountability frameworks; we return to this point below. Second, for many students with disabilities, the outcomes assessed within general education accountability systems have been viewed as irrelevant to the settings and skills required for successful post-school adjustments. Third and relatedly, the 1975 Individuals with Disabilities Education Act (IDEA), known as the Education for All Handicapped Children Act of 1975, required the formulation of an Individualized Education Program (IEP) for each student with a disability. This set the stage and provided legal endorsement for an individually-referenced and separate mechanism for describing the progress of students with disabilities.

Special education has, therefore, developed its own methods to account for its outcomes. Most special education accountability systems are focused exclusively on the

IEP, whereby student learning is measured relative to the individual student's goals. Such a framework can create technical and practical problems, while offering the advantage of individualizing outcomes. This advantage can be important when focusing on students whose expected outcomes vary substantially from the norm. In this section, we provide an overview of the IEP as an accountability mechanism. First, we describe the IEP document along with the typical assessment practices associated with the formulation and implementation of IEPs. Then, we summarize the problems associated with the historical implementation of IEPs within special education. Next, we provide a brief overview of alternative, more promising approaches for enhancing the use of IEPs in order to increase expectations and document student learning.

What is the IEP?

According to federal legislation, an IEP must be developed for every child in special education to define an appropriate education, to guide the delivery of services, and to frame methods for evaluating student outcomes. With respect to this last function, which is most pertinent to this document, the IEP must include a statement of the student's current levels of educational performance and a statement of measurable annual goals, including short-term objectives or benchmarks. Generally, an annual goal specifies the individual student's content and performance standards; it also structures the assessment standards by framing the end-of-year summative evaluation mechanism. The discrepancy between the current performance level in an area and the annual goal for that area indicates how "high" those standards are for an individual student. The short-term objectives or benchmarks create the framework for ongoing monitoring of student progress toward the accomplishment of the annual goal.

To identify the areas of need for which current performance levels, goals, and objectives must be specified, most IEP teams rely on commercial, individually-administered, norm-referenced tests of aptitudes and achievement. The normative scores provided by these measures permit the identification of areas in which a student's performance deviates from that of similar peers or from his/her own performance in domains (the typical cutoff for "deviation" is more than 2 standard deviations). Advantages of these measures include strong criterion validity, full-scale reliability, and assessment of performance over a wide range of levels. Problematic features of some

measures commonly used for these purposes include questionable construct validity, poor reliability of subtest scores (and, relatedly, intra-individual comparisons), and lack of utility for instructional program planning.

Because of the questionable utility of these measures for formulating relevant instructional programs, IEP teams often rely on alternative, more informal assessments when it comes time to identify current performance levels, goals, and short-term objectives or benchmarks within a demonstrated area of need. One of the most commonly used instruments for identifying current performance levels, goals, and short-term objectives or benchmarks is the Brigance Inventories.

To illustrate how this and other similar measures are used to formulate IEPs, we describe the application of the Brigance Diagnostic Inventory of Early Development (Brigance, 1978). This criterion-referenced assessment divides performance into 11 domains (pre-ambulatory motor, gross motor, and fine motor, self-help skills, pre-speech, speech and language, general knowledge and comprehension, readiness, reading, manuscript, and basic math). The tester, who according to the manual requires no special training, matches the area of need (which was identified through norm-referenced assessment) to one of these 11 domains. Items in a domain are organized hierarchically, suggesting a sequence in which normally developing children acquire skills represented by the items. Next to each item (except in reading and math), a year and month identifying typical acquisition guides the tester's choice about where to initiate assessment, and ceilings (i.e., number of failed items) guide termination of testing within a domain. Brigance directions indicate how to use performance to frame current performance level statements and goals: Current performance levels are the "skills of the highest level in the skill sequence; the objectives to be mastered are ... the skills immediately following those mastered ... are the logical skills to be developed during the next instructional period" (pp. vii; 252). Objective statements that correspond to each item are provided; however, no guidelines about how many objectives should be mastered during one school year or about how to relate annual goals to short-term objectives or benchmarks are provided.

Using this or other similar assessments, the IEP team might frame the current performance level for a particular area as, "The student, when provided with the

appropriate stimuli, says one word"; the goal as, "By May 31, the student, when provided with the appropriate stimuli, says 3-word phrases"; and the short-term objectives as, "When provided with the appropriate stimuli, the student will say three words other than mama or dada by October 31; use abbreviated statements by November 30; name two-or more familiar objects when asked their name by December 20; use phrases with adjectives such as big, good, little by January 31; use subject-predicate phrases by February 28; use plurals (adding s) by March 31; and use noun phrases by April 30" (see p. 124 of Brigance). As suggested in the manual, the special education teacher might use the Brigance to track mastery on each of these objectives and the goal; the Brigance includes one item to assess each skill. In fact, special educators often rely on commercial or teacher-made criterion-referenced measurements to track attainment of IEP goals and objectives; if they do, decisions about how frequently to measure and whether to use alternate forms are formulated idiosyncratically. Most commonly, however, teachers rely on informal observations of student performance during instructional sessions to index mastery (Potter & Mirkin, 1982).

Problems with Typical IEP Use

Problems with typical IEP use might be categorized into three areas: procedural errors, reliance on faulty assessment methods, and substantive problems.

With respect to procedural errors, research implicates typical IEP practice as incomplete and faulty. In analyses of existing IEP documents, for example, Smith and Simpson (1989) reported low numbers of completely stated goals and substantial inconsistencies between current performance statements and annual goals. Smith (1990) also documented how teachers did not record objectives as they were attained.

In terms of reliance on faulty assessment methods, because the assessment tools used to specify goals and monitor progress toward goal attainment often are criterion-referenced, they typically have unknown psychometric properties, raising concerns about the meaningfulness and accuracy of the database on which goals and goal attainment are based. Moreover, most criterion-referenced measures focus narrowly on discrete, decontextualized skills, making the formulation of broadly generalizable, long-term goal statements difficult. As stated in the Brigance manual, for example, to develop "the skill

sequences and developmental ages, many professional materials ... [were] examined and cited" (p. iv); no empirical basis for reliability and validity of the measure is provided. Moreover, according to the manual, "tests results are valid [even] if assessment procedures are not followed rigidly or ... adaptations are made" (p. xi). Such statements reflect a disappointingly nonempirical approach to reliability and validity often reflected in criterion-referenced instrumentation (Tindal, Fuchs, Fuchs, Shinn, Deno, & Germann, 1985).

Aside from technical errors in the manner IEPs are implemented and the assessment tools upon which they are based, it is disturbing to find that IEP use typically does not conform to the substantive spirit reflecting federal legislation. Rather, IEPs have served primarily as a tool for procedural compliance monitoring, whereby federal auditors make sure that a complete IEP exists for each student receiving special education services and that IEPs document how (i.e., where, when, and by whom) those services are delivered (Smith, 1990). Research suggests that IEPs are not frequently used as a guide for framing high expectations for students with disabilities or providing documentation for how much and what pupils have learned (Wesson, Deno, & Mirkin, 1982).

Enhancing IEP Use for Documenting Student Learning

Despite these serious problems associated with conventional IEP practice, the original assumption among the framers of the original IDEA was that IEPs should provide a structure for setting high standards and measuring student outcomes. This perspective is reflected broadly in the special education literature (e.g., Fuchs, Fuchs, & Hamlett, 1990) and represented in the 1997 amendments to IDEA.

The question is, do methods exist to reorient the IEP process toward addressing substantive, in addition to procedural, compliance so that IEPs provide a framework for increasing expectations and monitoring student outcomes? In fact, well-developed methods do exist for accomplishing these functions and, in some states, are practiced widely.

Within such frameworks, the IEP does not, in contrast to conventional practice, specify the numerous subskills a teacher might plan to teach sequentially during a school year and does not rely on weak assessment methods. Instead, the IEP identifies the broad

outcomes, along with validated indicators of proficiency on those outcomes, that the student is expected to perform by the end of the year. Research demonstrates that these alternative frameworks can result in more ambitious goals for students with disabilities (e.g., Fuchs, Fuchs, & Hamlett, 1990) as well as stronger student learning (e.g., Fuchs, Fuchs, Hamlett, & Stecker, 1991; Wesson, 1991).

Therefore, available methods more closely reflect the assumptions embedded within current, standards-based education reform: reorienting practitioners toward a stronger focus on student outcomes and high standards. In contrast to standards-based reform, however, these alternative structures permit consideration of individual goals for students whose goals do not correspond well to standards-based reform's focus on challenging cognitive content.

Nevertheless, within such an individually-oriented outcomes framework, technical problems remain in aggregating information across students. Moreover, the difficulty associated with implementing a professional development agenda necessary to retool special educators toward a reoriented IEP process, which is designed to increase expectations and measure meaningful outcomes, cannot be underestimated. In fact, such a professional development agenda parallels the task of reorienting the general education community to the high standards and outcomes orientation of the standards-based reform movement.

Participation in General Education Accountability Frameworks

Nearly every state and many school districts and schools now have some kind of accountability framework in place (Bond & Roeber, 1995). In 1998, 49 states had active statewide assessment programs; only one state had no plans to implement a statewide assessment program of any kind.

Several studies have documented that the participation of students with disabilities in statewide assessments historically has been minimal, with extremely variable participation from one state to another (Erickson, Thurlow, & Thor, 1995; McGrew, Thurlow, Shriner, & Spiegel, 1992; Shriner & Thurlow, 1992). The low participation rates of students with disabilities has been documented despite (a) the difficulty of calculating comparable figures across locations and (b) the tendency of states to calculate participation rates in ways that inflate estimates (Erickson, Thurlow, & Ysseldyke, 1996).

Decisions about whether to include students with disabilities in general education accountability frameworks typically are formulated by IEP teams (Erickson & Thurlow, 1996). Several factors have been shown to contribute to IEP team decisions, including (a) the use of ambiguous decision-making guidelines, which often focus on superficial considerations rather than the educational goals and learning characteristics of students; (b) concern about the potentially negative emotional impact of forcing low-performing students to complete challenging assessments (Ysseldyke, Thurlow, McGrew, & Shriner, 1994); and (c) extra-student variables that pressure schools to exclude low-performing students in order to inflate the aggregated data.

Moreover, in many states, a student's participation in statewide assessment does not necessarily mean that the student is included in public accountability reports. Most commonly, these students' scores are flagged and excluded from data aggregation simply because the students have identified disabilities (see Thurlow, Scott, & Ysseldyke, 1995b). In addition, scores of students with disabilities frequently are excluded when state accountability data are aggregated because of concern that test accommodations may invalidate scores (Thurlow, Scott, & Ysseldyke, 1995a). Below, we focus additional attention on this last concern: whether test accommodations invalidate scores.

Initiatives to Broaden Public Accountability Frameworks

In light of (a) the current demand for accountability about student outcomes, (b) the problems documented with a separate special education accountability system, and (c) historically low participation rates for students with disabilities within general education frameworks, urgent concern exists about the participation of students with disabilities in state accountability programs. As a reflection of this concern, IDEA 1997 requires states and districts to include students with disabilities in their state- and district-wide assessment programs. The assumption is that if schools are to consider the needs of students with disabilities deliberately and proactively in reform and improvement activities, the outcomes of students with disabilities must be represented in public accountability systems (McDonnell, McLaughlin, & Morrison, 1997).

As already mentioned, one major reason for excluding students with disabilities from public accountability systems is that no widely agreed upon methods exist for determining which test accommodations preserve the meaningfulness of scores for which students with

disabilities (McDonnell et al., 1997). Accommodations are changes in standardized assessment conditions introduced to "level the playing field" for students by removing the construct-irrelevant variance created by their disabilities. Valid accommodations produce scores for students with disabilities that measure the same attributes as standard assessments measured in nondisabled individuals. On the one hand, disallowing valid accommodations prevents students with disabilities from demonstrating their abilities. On the other hand, overly permissive accommodation policies inflate scores and inadvertently reduce pressure on schools to increase expectations and outcomes for students with disabilities (McDonnell et al., 1997).

Lack of consensus about appropriate accommodations is revealed in variations among state policies; in fact, some states prohibit accommodations that other states recommend (Thurlow, Erickson, Spicuzza, Vieburg, & Ruhland, 1996). Moreover, decisions for individual students with disabilities typically are formulated idiosyncratically by IEP teams (Erickson & Thurlow, 1996), with vague decision-making rules that often focus on superficial variables (Ysseldyke, Thurlow, McGrew, & Shriner, 1994). Without well agreed upon criteria for determining which accommodations, if any, are valid, comparisons between states or districts are not meaningful. In response, states and districts often exclude students with disabilities altogether, or exclude students who have been tested with accommodations from public reports, or disallow the use of all accommodations that violate standardized testing conditions.

Clearly, school personnel require information about the effects of test accommodations on students with disabilities. With clear information, schools might arrive at empirically-based conclusions about which accommodations are allowable for which students with disabilities. With consistent decision-making criteria across schools, districts, and states, a "level playing field" might be achieved not only for students with disabilities but also for school, district, and state comparisons -- with all students with disabilities included in the database.

Purpose of This Document

With this goal in mind, our purpose in preparing this document is to provide school district and state department personnel with a comprehensive synthesis of the research literature on the effects of test accommodations on students with disabilities. To provide

readers with a framework for interpreting this research synthesis, we briefly explain and illustrate Tindal's (1998c) recent classification of research approaches for examining the validity of test accommodations. According to Tindal, research approaches on this issue may be classified as descriptive, comparative, or experimental.

The most common approach to determining the validity of test accommodations is descriptive. With a descriptive approach, accommodations are analyzed logically to consider the nature and severity of the disability the accommodation will offset, along with the characteristics of the assessment.

For example, consider a student with a visual disability who takes a mathematics concepts and applications test that requires text reading in problem stimuli. One might logically conclude that a large-print accommodation is valid because it permits this student to access the printed information and thereby allows the student to demonstrate his or her mathematics competence while preserving the meaningfulness of the measured construct. According to Phillips (1994), one important indicator that an accommodation serves to level the playing field between students with and without disabilities is differential boost. That is, the accommodation increases the performance of students with disabilities more than it increases the scores of students without disabilities. For the large-print accommodation, logical analysis dictates that while the accommodation makes the assessment more accessible to students with visual disabilities, the accommodation slows down students without visual disabilities.

Of course, for some populations, logical analysis of accommodations is more difficult. This is the case for students with learning disabilities (LD), who constitute more than half of the students with disabilities. Logical analysis of test accommodations for students with LD is challenging because the LD population is heterogeneous. This makes conceptual analysis of meaningful accommodations impossible, as it might be for visual disabilities, and it dictates empirical study with a strong focus on individual differences (McDonnell et al., 1997). The second problem is the nature of the cognitive problems students with LD present. The most distinguishing characteristic of students with LD is reading and math deficits (Kavale & Reece, 1992) -- while most high-stakes assessments directly measure or rely heavily on reading and math skills. Therefore, many accommodations currently used to address the disadvantages inherent in the LD

population (e.g., extended time, decoding questions, encoding responses) may distort the meaning and interpretation of scores. Because the disability is intertwined with the constructs measured, allowing accommodations may effectively exempt students with LD from demonstrating the cognitive skills the test measures (Phillips, 1994). Due to these difficulties, additional approaches, beyond descriptive or logical analysis, are necessary for students with LD.

With a comparative approach to studying the effects of accommodations, extant databases are analyzed to gain insight into how accommodations may affect students with disabilities. For example, the Educational Testing Service has examined scores of students with and without disabilities to compare performance with and without accommodations (see Willingham, 1989). Findings indicate that special and regular administrations of the SAT are comparable, with the exception of the extended time accommodation -- probably the most common accommodation for students with LD. Of course, the population of students with disabilities taking the SAT is not broadly representative.

To examine a more representative sample using a comparative approach, Koretz (1997) retrospectively analyzed the performance of Kentucky students who received dictation, oral reading, rephrasing, and cueing accommodations on the Kentucky assessment. Koretz identified disturbing patterns in the data, which indicated that accommodations overestimated the academic competence of students with disabilities. Findings raised questions about how the accommodations were administered.

As demonstrated with these studies, a comparative approach to determining test accommodations permits interesting insights into the effects of accommodations. Nevertheless, the retrospective analyses of extant databases inherent with a comparative approach often leaves important questions unanswered. Moreover, the comparative approach does not advance understanding of whether accommodations produce differential boosts for students with disabilities over and above what we might expect for students without disabilities. This question, of course, is essential to issues of validity (Phillips, 1994). To provide this information, an experimental approach is necessary.

With an experimental approach, the effects of accommodations are examined with controlled, prospective research designs, which examine effects for students with and without disabilities when tests are administered with and without accommodations. In one

of the most carefully controlled study to date, Tindal, Heath, Hollenbeck, Almond, and Harniss (1998) reported analyses for students with disabilities and without disabilities in reading and mathematics. In reading, students completed half the test by bubbling in responses in standard fashion; the other half, by marking responses directly on the test booklet. In math, students either read the math test silently to themselves or listened as the teacher read the test aloud. Results indicated that the response accommodation (i.e., bubbling an answer sheet vs. marking directly on the test booklet) did not affect students' reading scores. By contrast, the math accommodation differentially affected students with disabilities: Scores with the reading aloud accommodation improved statistically significantly more for students with disabilities than for students without disabilities. This suggests the validity of a reading aloud accommodation for mathematics tests for students with disabilities.

Overview of This Research Synthesis

In this research synthesis on the effects of test changes, we assembled the research to date on test accommodations. Because this field of inquiry is expanding so rapidly, it is likely to serve three primary purposes. First, the structure of the review and the manner in which information is organized may serve as a model for organizing current and future research. We used a taxonomy from the National Center on Educational Outcomes (NCEO), an organization that has been studying state assessment policy for the past decade. Second, the information itself can be used directly to justify the use of certain accommodations. For example, as noted above, in many states the use of accommodations lacks an empirical basis. To the degree that the assessment information is used to make a high stakes decision and data are available, they should be referenced. This document would be used to help personnel responsible for testing and special education make such references. Third and last, the outcomes from this summary should be used to reflect on the manner in which the research is generated, the quality of the findings, and the conclusions we make as a field. Some of the research we summarize is excellent in methodology, and conclusions can be made without qualification. Other research we summarize needs to be qualified, often because of the sheer difficulty in providing rigorous experimental controls, leaving the researchers to rely on quasi-experimental strategies instead. To serve these three purposes and to ensure the broadest foundation possible, we began with a description of our

methodology for searching the literature, then described the organization of the research summary, and finally concluded with a reflection on the truthfulness or validity of the findings. As will be described later, we used several criteria in determining whether studies would be included in this review.

Search for and Organization of Test Changes Research

In our identification of the literature on test changes, three major databases were searched: (a) ERIC, (b) PsychInfo, and (c) Dissertations Abstract International (DAI). A fourth search was added after the fact by using the work of Chiu and Pearson (1998), who completed a very thorough analysis of this last database and published the abstracts from these dissertations on a web site (http://pilot.msu.edu/~chiuwing/Dissertations.htm), which was then analyzed in our review. Because most of these dissertations were ordered but unavailable, we wrote directly from these abstracts. Table 1 is a list of the word search used for each of these databases, it provides the number of references appearing for each word in each database.

Because key word searches are literal, terms were varied systematically to include all meaning associated with a general sense of changes made in testing. We were most interested in large scale testing and refereed research so we did not follow up on the 1,161 references to *state test changes* listed in Dissertations Abstract International (DAI), which is noted with an asterisk(*). Also, in this search, we focused on understanding the effects of these changes on students with disabilities (SWD) though we included studies if they were conducted in a category listed by NCEO (see page 16) and had implications for SWD.

Table 1. Search Terms and Sources for Identifying Research on Test Accommodations

Number of Citations Found Using Different Key Words/Phrases			
Words Used	ERIC	Psychinfo	DAI
Test changes standardized testing	15	0	43
Test changes large-scale testing	0	0	3
Large-scale test changes	0	0	36
Standardized test changes	29	0	263
Test changes state testing	5	0	134
State test changes*	5	0	1,161
Test changes standards-based testing	0	0	0
Standards-based test changes	0	0	0
Standardized test modification	5	0	35
Large-scale test modifications	0	0	9

Table 1. cont.

Words Used	ERIC	Psychinfo	DAI
State test modifications	2	0	110
Standards based test modifications	1	0	6
Alternate assessment standardized	0	0	18
Alternate assessment large-scale	0	0	3
Alternate assessment state-testing	0	0	4
Alternate assessment standards-based	0	0	4
Alternative assessment standardized	15	0	85
Alternative assessment large scale	1	0	16
Alternative assessment state-testing	1	0	13
Alternative assessment standards-based	0	0	25
Standardized testing accommodations	3	0	1
Large-scale testing accommodations	0	0	0
State testing accommodations	3	0	7
Standards-based testing accommodations	1	0	0
Test changes		70	
Test modifications		20	
Alternate and alternative assessment		112	
Test accommodations		5	

In addition to this computer search, we conducted a hand search of current measurement and special education journals. We looked at tables of contents for the following journals and years, looking for articles using our general key words above.

Applied Measurement in Education (vol. 3, 1990 to vol. 12, 1998)

B.C. Journal of Special Education (1982 to 1998, Vol 3)

British Journal of Special Services (1985 to March 1998)

Diagnostique (1993 (vol 18, no 2) to 1997 (vol 23, no 1))

Educational Assessment (vol 1, 1993 to vol. 5, 1998)

Educational and Psychological Measurement (1976 to Aug 1998)

Educational Measurement: Issues and Practice (1982 to Summer 1998)

Educational Psychologist (1976 to Winter 1998; except 1986 No. 4)

Educational Psychology (1981 to June 1998)

Educational Researcher (1976 to May 1998)

Exceptional Children (1976 to Summer 1998)

Journal of Educational Measurement (1976 to Spring 1998)

Journal of Learning Disabilities (1976 to July/Aug 1998; except 1991, 1992, 1981)

Journal of School Psychology (1976 to Summer 1998; except vols. 19, 20)

Journal of Special Education (1976 to Summer 1998; except 1980, 1976)

Learning Disability Quarterly (1982 to Summer 1998; except 1988, 1989)

Remedial and Special Education (1984 to July/Aug 1998), was previously Exceptional Education Quarterly (1980 to 1984)

School Psychology Review (1976 to 1998, vol. 2; except 1990 No. 4 and 1994 No. 4)

Finally, we followed secondary references back from each primary article we found in either the computer or hand search. Along with the bibliography assembled from our search, we collected the references from the earlier work by NCEO in their two publications (Thurlow, Ysseldyke, & Silverstein, 1993; Thurlow, Hurley, Spicuzza, & Sawaf, 1996).

Research Review Criteria

In our search we concentrated on large-scale testing, whether or not the actual study included a large-scale test. Probably the most important criterion was that the test had to reflect a broad measure of achievement and not reflect a criterion test used as part of classroom instruction. This latter literature was deemed inappropriate primarily because such testing often is used to evaluate the effects of specific curriculum or instructional strategies rather than document overall levels of achievement. Generally such large-scale tests are group administered. We included, however, individually administered tests because such administration changes frequently are allowed in state policies. Generally such tests are achievement and not ability measures. We included both types, however, because of the high correlation between them. Although we required the change in testing to provide outcomes (empirical or experimental) we did not use experimental integrity as a criterion and included all studies in which a change in testing was documented. As a result, all conceptual or policy manuscripts or articles were ignored. Finally, we primarily were interested in testing students with disabilities but we included some studies on students without disabilities, particularly if the previous criteria were fulfilled and the outcomes had bearing on test changes for students with disabilities.

Research Organization Criteria

The National Center on Educational Outcomes at the University of Minnesota has been tracking state policies and practices on statewide testing for nearly a decade. Assessment change and adaptations range from allowing no modifications to permitting specific

accommodations for some students. Like the issue of inclusion, changes in testing and measurement practices are rife with controversy. Generally test changes are grouped into two types: (a) accommodations , and (b) modifications. Accommodations are considered changes in the way the test is given or taken but do not alter the central construct being measured by the test. In contrast, modifications are considered substantial changes in the way the test is given or taken and definitely alter the construct being measured by the test. A major purpose of this document is to ascertain any empirical support for this distinction. In this review, accommodations and modifications have been grouped into the following four categories, using the structure from Ysseldyke, Thurlow, McGrew, and Shriner (1994):

- •Presentation adaptations in which stimuli (materials) presented to students are modified.
- •Response changes with students allowed to use a different manner of responding.
- •Setting adaptations in which variations are made in the context of where tests are administered and who administers the test.
- •Timing and scheduling adaptations in which changes are made in how long and how many sessions a test is administered.

These changes have been listed, in Table 2 with **those in bold being reviewed** in the following sections of this paper, both in text and table form. We were unable to locate a sufficient research base to report on the accommodations in plain text. Because of the increase in this research, however, it is very likely that new research is forthcoming.

Table 2. List of Test Changes With (bold) and Without Research (normal)

Timing/Scheduling	*Setting*
•Use flexible schedule	**•Administer the test individually in a separate location**
•Allow frequent breaks during testing	**•Administer the test to a small group in a separate location**
•Extend the time allotted to complete the test	**•In a small group, study carrel**
•Administer the test in several sessions, specify duration	•Provide adaptive or special furniture
•Provide special lighting	•Administer test in locations with minimal distractions
•Time of day	
•Administer test over several days, specify duration	
•Provide special acoustics	

Presentation	*Response*
•Braille edition or large-type edition	<u>Test Format</u>
•Prompts available on tape	•Increase spacing
•Increase spacing between items or **reduce items**/page-line	•Wider lines and/or wider margins
•Increase size of answer bubbles	•Graph paper
•Reading passages with one complete sentence/line	•Paper in alternative format (word processed, Braille, etc.)
•Multi-choice, answers follow questions down bubbles to right	**•Allow student to mark responses in booklet instead of answer sheet**
•Omit questions which cannot be revised, prorate credit	
•Teacher helps student understand prompt	
•Student can ask for clarification	<u>Assistive Devices/Supports</u>
•Computer reads paper to student	**•Word processor**
• Highlight key words/phrases in directions	•Student tapes response for later verbatim transcription
	•Calculator, arithmetic tables
<u>Test Directions</u>	**•Spelling dictionary or spell check**
•Typewriter	**•Alternative response such as oral, sign, typed, pointing**
•Dictation to a proctor/scribe	•Brailler
•Communication device	•Large diameter, special grip pencil
•Signing directions to students	•Copy assistance between drafts
•Read directions to student	•Slantboard or wedge
•Reread directions for each page of questions	**•Tape recorder**
•Simplify language in directions or problems	•Abacus
•Highlight verbs in instructions by underlining	•Provide additional examples
•Clarify directions	
•Provide cues on answer form	
<u>Assistive Devices/Supports</u>	
•Visual magnification devices	
•Templates to reduce visible print	
•Auditory amplification device, hearing aid or noise buffers	
•Audiotaped administration of sections	
•Secure papers to work area with tape/magnets	
•Questions read aloud to student	
•Masks or markers to maintain place	
•Questions signed to pupil	
•Dark heavy or raised lines or pencil grips	
•Assistive devices-speech synthesis	
•Amanuenis (scribe)	

To organize our summary, the test changes have been listed in a separate table as a histogram with each study numbered (see Table 3). For each change that was the main focus of research, the histogram simply lists the study number, which is the ordinal number of the study in the reference list, therefore reflecting an alphabetic order. The table presents changes in the order of frequency in which studies have been done within an accommodation area. Some studies are included in more than one change (for example, in the research on the Kentucky state test, KIRIS, the effects from several different changes were analyzed, including dictation, cueing, rephrasing, and oral reading). Finally, research is included in the table with and without students with disabilities being part of their subject sampling. For those studies with no students with disabilities, the number in the histogram has been shaded.

Table 3. Test Changes in which Research is Summarized

TIME/SCHEDULING

Extended time to complete test	2	*6*	8	15	21	33	34	36	*38*	44	50	57
	59	*67*	68	69	70	*71*	72	76	79	82	94	108
	115											

SETTING

Separate location/Small group	*21*	91
Auditory stimulation	1	

PRESENTATION-Single Focus Areas

Computer presentation	14	*18*	*39*	40	*45*	48	*52*	*55*	*56*	*61*	65	*90*
	93	*104*	*106*	107								
Examiner familiarity	*21*	25	26	27	*28*	29	30	32	74	91		

PRESENTATION-Multiple Changes Areas

Large type/Braille	7	8	9	17	35	64	75	82
Read items on test (rephrase or cue)	33	37	54	94	100	102	110	
Pace/Reduce items per page/Video present	18	41	45	99				
Audio taped administration	9	24	75	98				
Change answer sheet/Cues on test/Reduce distractions	7	35	49	77	105			
Levels of syntax	66	111						
Time, separate testing, reading, alternate response, Braille	73	114						

RESPONSE

Dictation to a proctor/scribe	33	42	43	54	62	102	
Alternate assessment/ Alternative response	4	10	19	20	22	23	92
Mark responses in test booklet	64	83	100				
Work collaboratively with other students	31	78	87	109			

ASSISTIVE DEVICES/SUPPORTS

Word processor	5	40	43	46	47	62	80	81	101	103
Calculator	13	16	33	60						

OTHER

Reinforcement	11	12	49	51	53	86	89	95	96	97	113	114
Instruction on test taking strategies	63	84	85	88	112							
Instructional level testing	58											

A REVIEW OF WHAT WE KNOW SO FAR

In this section of the manuscript we present a brief summary of the research on the test changes noted in Table 3. In reviewing this research, two different formats are employed. First, the research is summarized in text, with attention to analysis of literature by subjects and test and by research quality and summary. Generally, commonalities across the studies are highlighted. Second, the research is summarized in tables that include information on the authors [and study number], year of the study, description of the test change, subjects included in the study, test given as the dependent variable, and the findings reported from the research. In these tables, the research is listed in alphabetical order within each accommodation area. Studies included in more than one section are presented in tabular format only once.

Timing/Scheduling of Testing

In this category, we have included the amount of time a student is allowed to complete the test. In the literature, this issue frequently appears with terms like *speediness* or *timed testing*. As Gulliksen (1950) noted long ago, mental tests are a combination of both speed and power. Speed deals with the degree to which performance is measured under timed or untimed conditions. Power addresses the difficulty or depth-breadth of the item sampling plan. The typical changes in this category involves giving the student more than the designated times and/or allowing the test to be taken in several sessions or in briefer durations.

One of the earliest studies to investigate speed as a unique dimension of tests was conducted by Baxter (1931). For him, the crucial issue was the length of time it took the student to complete the test using items from all levels of difficulty. He reported very

high relationships between speed and level [of difficulty] and a significantly greater contribution by speed in determining power over level. Meyers (1952) also completed similar research 20 years later using different measures. In this later study, he operationalized speed as "the percentage of persons marking responses at various points through the test" (p. 349) and concluded that the score on speeded tests is a function of two unrotated (orthogonal) factors, ability and rate-of-answering, with ability being more valid. Lord (1956) continued this line of research with a related study of the influence of speed factors in ordinary aptitude tests. Using admissions examinations, grades, and several experimental vocabulary, and arithmetic measures that varied on speediness, he identified number-speed, verbal-speed, and spatial speed factors, and found they were highly intercorrelated. Finally, Mollenkopf (1960) reported higher correlations between speeded and nonspeeded versions of verbal tests than for arithmetic reasoning tests. Some remarkable changes in the rank ordering also were noted for some students on the math test when they were allowed enough time to finish (e.g., changing from the 63rd to the 12th from the top or the 68th to the 18th from the top). "The added time apparently allowed these students to better show their stuff, and to come up appreciably in their standings" (p. 226). Research also has been done by Evans and Reilly (1972a, 1972b, 1973) and Evans (1980). Their work typically has focused on the effect of increasing or reducing time for students of different ethnic backgrounds. Virtually all of this work has been on older students, either of college or high school age, and using admission tests or entrance examinations.

Analysis of Literature by Subjects and Test

Most of the research on speededness has been completed with college age students. In time sequence, the research has been done by Hill (1984), Centra (1986), Tachibana (1986), Bennett, Rock, and Jirele (1987), Rock, Bennett, and Jirele (1988), Halla (1988), Linder (1989), Derr-Minneci (1990), Weaver (1993), Powers and Fowles (1996), Ziomek and Andrews (1996), Alster (1997), Jensen (1997), and Ofiesh (1997). Of course, within the college age range, the type of testing investigated has tended to be oriented toward either admissions tests (such as the Scholastic Aptitude Test, American College Test, or the Graduate Record Examination) or skill tests (such as the Nelson-Denny, (Brown, Fishco, & Hanna, 1981-1993). When studying the impact on students with disabilities,

the largest category sampled has been students with learning disabilities. Significant effects typically have been reported, mostly in terms of the relationship between extended time and performance on the admissions tests or later grade point averages obtained while in college. On occasion, positive effects have been large (see Centra, 1986). See the section on *Presentation and Response - Multiple Changes* for further descriptions of research by ACT and ETS.

A few studies have begun to appear on the effects of timed testing on students from the Kindergarten-grade 12 population on students with varying disabilities and achievement levels. Murray (1987) studied middle school students taking a spatial relations test and found untimed tests benefited students with learning disabilities who also were of average skill in math. Likewise, Perlman, Borger, Collins, Elenbogen, and Wood (1996) studied timed testing with middle school students using the Iowa Test of Basic Skills (ITBS) and reported positive outcomes from providing extended time. Gallina (1989) compared timed and untimed tests for students with Tourette's Syndrome (TS) or Attention Deficit Hyperactivity Disorder (ADHD) and found positive outcomes for students with TS only on the Wide Range Achievement Test (but not on the Metropolitan Achievement Test). For Munger and Lloyd (1991), no differences were found for 5th grade students who took the ITBS, whether they were with or without disabilities. This finding of no effects also was reported by Harris (1992) with 16 high school juniors who took the Preliminary Scholastic Test (PSAT) in timed and untimed conditions. With a population of "young children," Montani (1995) found scores on untimed story problem and computation tests to be higher than timed tests if the student had difficulties in math but not reading (who performed comparable to students without reading or math difficulties, in both the timed and untimed conditions). On a population of elementary students taking an oral reading fluency measure, Derr-Minneci (1990) reported higher performance when the test was timed. This finding was the only positive outcome for the use of timed tests; however, it reflects the only measure in which rate of behavior is the critical datum and therefore renders the finding less unusual. Finally, for Fuchs, Fuchs, Eaton, Hamlett, and Karns (in press), when students with learning disabilities were allowed to take conventional math tests (computations and concepts/ applications) with extended time, they did not benefit more than students without learning

disabilities. On more innovative tests of extended math problem solving, however, students did benefit differentially from extended time.

Analysis of Research Quality and Summary

Much of the research with college age students and using admissions tests or entrance examinations has relied on a quasi-experimental research design. Students with or without disabilities have taken part in both conditions of test administration (timed and untimed), and then a one-between, one-within ANOVA has been conducted. In most studies, the effect of allowing students to take the test untimed was positive. Sometimes this finding occurred only for students with disabilities, although in some studies, removing time restrictions benefited all students. The studies completed with the K-12 population have included students from many different disability groups and have used a broad range of published achievement tests, as well as generic proficiency and skill measures. These studies also reflect varied research methodologies: Some designs have been less robust in their sampling of students or analyses of data, rendering the findings less interpretable. For example, Harris (1992) sampled too few students; Perlman et. al. (1996) had difficulties in obtaining subjects; and Jensen (1997) incorrectly analyzed percentile rank data.

In conclusion, if the purpose of the test is to gain information on broad band achievement measures, timed conditions may not allow students with disabilities to reflect their full "abilities" and may actually introduce error variance. This conclusion is at best uncertain, however, because most research fails to compare effects for students with and without disabilities. Some well-controlled work (comparing students with and without learning disorders) indicates that on conventional math tests, extended time does not differentially benefit students with disabilities. Additional research on this frequently used test accommodation is needed. This research needs to be explored to identify cause and effect relationships in determining whether extended time is needed because of reading problems, information processing, or any number of reasons that appear to be disability-related.

Annotated References of Investigations on Timing/Scheduling

Alster [2] 1997

Adaptation The tests were given under two conditions: Timed and Untimed.

Subjects Participants included 88 students, 44 with LD and 44 without disabilities, from five California suburban community colleges. The mean age for the students with disabilities was 26.7 and the mean age for the students without disabilities was 25.3. Each group consisted of 27 females and 17 males. About 57% of the participants were Caucasian, 32% were Mexican American, 7% were Asian, 2% were African American, and 2% were Native American. For 6 of the students with LD and 8 of the students without disabilities English was a second language (ESL).

Dependent Variable The ASSET Elementary Algebra Test (American College Testing Program [ACT], 1989) was used. For this study, the 25 problem test was divided by type of problem and level of difficulty into two comparable tests. One of the problems was deleted so each of the forms had 12 problems.

Findings The students with LD scored significantly lower than students without disabilities on algebra tests under timed conditions. The untimed scores of students with LD, however, did not differ significantly from timed or untimed scores of students without disabilities.

Baxter [6] 1931

Adaptation The test was given in less time (16 minutes) instead of 21 minutes. Three variables were analyzed: Speed, Level, Power

Subjects College students with no specific description other than they were in R.O.T.C.

Dependent Variable The main dependent measure was the Otis Self-Administering Test. Three criterion variables were used:

- The Army Alpha Revised Form Five with the following measures collected: number of items correct in limited time and unlimited time and the time taken to complete the test
- The College Aptitude Test
- Honor point ratio

Findings For the College Aptitude Test, level had greater validity with power. Power was found to have greater validity with the criterion of grades than speed or level. Speed and level in combination provide a significant contribution to power. Speed and level are independent. When measured in groups, the validity of level and power decreases.

Centra [15] 1986

Adaptation The tests were given under two conditions: Timed and Untimed

Subjects Approximately 1800 students, 79% with learning disabilities.

Dependent Variable The dependent measure was the Scholastic Aptitude Test (SAT).

Findings Students with LD improved their performance with extended time, the increase being greater than for students without disabilities tested with extra time. The average gains over scores earned in a timed administration were generally between 30 and 38 points on the SAT after growth in student ability, practice effects, and error of measurement were taken into account. About one in seven students gained over 100 points; between 3 and 7 percent decreased by at least 50 points. Score gains increased as time spent on the test increased, suggesting that the additional time was needed to reduce the effects of the examinee's disability.

Fuchs, Fuchs, Eaton, Hamlett, & Karns [33] in press

Adaptation Two computation curriculum-based measures (CBMs) were administered: one in standard fashion and one with extended time. Four concept and application CBMs were administered in standard fashion, with extended time, with calculators, and with reading the text aloud to students. Five problem-solving CBMs were administered: one in standard fashion, one with extended time, one with calculators, one with reading text aloud to students, and one with encoding (i.e., writing responses for students, as requested). A large-scale assessment was administered under standard and accommodated conditions.

Subjects The study included 373 fourth graders. Approximately half of these students had no identified disability; the other half had a learning disability (LD).

Dependent Variable Brief CBMs were administered in three mathematics domains: computations, concepts and applications, and problem solving. A large-scale assessment, also called a multi-faceted assessment of mathematics problem solving, was given.

Findings On conventional tests of mathematics achievement, students with LD did not benefit more than students without LD from an extended time accommodation (for the computation and concept and application CBMs) or from a calculator or reading accommodation (for the concepts and application CBMs). In fact, with extended time and calculator accommodations, students without LD, on average, profited more than students with LD. On the more innovative problem-solving CBMs, students with LD benefited more than students without LD from three of four types of accommodations: extended time, reading, and encoding. Effects for the calculator accommodation were marginally significant.

Gallina [34] 1989

Adaptation The test was given under two conditions: Timed, Untimed.

Subjects Eighty-one elementary students were assessed: 27 students with Gilles de la Tourette's Syndrome (TS), 27 control students, 27 students with Attention Deficit Hyperactivity Disorder (ADHD).

ependent Variable Three assessments were used:

- Wide Range Achievement Test-Revised (WRAT-R), Arithmetic subtest
- Metropolitan Achievement Test (MAT6), Mathematics subtest
- ADD-H Comprehensive Teacher Rating Scale (ACTeRS), for evaluating social behavior

Findings Subjects with Tourette's Syndrome performed poorly on the WRAT-R arithmetic subtest and on the MAT6 under timed conditions but scored in the average range in the untimed condition. Students with Tourette's Syndrome have good social skills and little oppositional behavior when compared to students with ADHD, according to the ACTeRS.

Halla [36] 1988

Adaptation The tests were given under two conditions: Timed, Untimed.

Subjects One-hundred twenty-six subjects with and without LD (72 female, 54 male) participated in this study. Subjects were undergraduate, graduate, and non-students, ranging in age from 20 to 56.

Dependent Variable Two tests were used: Graduate Record Examination (GRE) General Test, Nelson-Denny Reading Test.

Findings Results indicated no significant effect on test performance for both students with and without LD. Timed scores for the students with LD were significantly lower than for the students without LD. Also, test scores increased significantly for both groups between timed and untimed testing conditions.

Harris [38] 1992

Adaptation The test was given under three conditions: timed and verbalized,untimed and verbalized, untimed and solved silently.

Subjects Sixteen high school juniors participated in the study. The students were classified on ability level as measured by scores on the Preliminary Scholastic Aptitude Test (PSAT).

Dependent Variable The assessment consisted of three sets of verbal and quantitative questions obtained from the PSAT.

Findings Upper level students and females employed more effective test taking behaviors and outperformed their counterparts on the actual test questions. Upper level black students outscored all other groups on both the skills measure and the test questions. Upper level students outperformed lower level students on the Verbal Skills measure and on the verbal test questions. Upper level black students attained the highest average score on Verbal Skills as well as on the verbal problem sets. On the math skills measure, upper level students outscored the lower level students. Verbalization during problem solving for verbal questions was not significant. However, thinking out loud while solving math problems did have a significant impact. Lower level students, especially lower level females, benefited from verbalizing during problem solving.

Hill [44] 1984

Adaptation The tests were given under two conditions: Timed and Untimed.

Subjects Ninety-six undergraduate college students participated: 48 students with LD and 48 students without LD. The two groups were matched in several areas including: Gender (25 males, 23 females), Race (47 Caucasians, 1 Hispanic).

Dependent Variable Two tests were administered: American College Test (ACT), Nelson-Denny Reading Test (1981 edition).

Findings Testing condition had a significant effect on test performance, but primarily for students with LD. In the timed testing condition, the ACT and Nelson-Denny raw scores for students with LD were significantly lower than the scores of students without LD. However, there was no difference between the two groups' test scores in the untimed condition.

Jensen [50] 1997

Adaptation The tests were given under two conditions: Timed and Untimed.

Subjects A total of 22 college students participated: 12 students without LD, 10 students with LD.

Dependent Variable A computerized version of the comprehension subtest in the Nelson-Denny Reading Test was used (forms G and H). The computerized test records the students reading rates, response times to questions, and question answers.

Findings The results show that there was a significant difference between students with LD and students without LD on the timed test, regardless of order of presentation. This difference was also found between students with LD and students without LD if the untimed test was presented second. However, there was no significant difference between the students with LD and stunts without LD if the untimed test was administered first. Additionally, reading rates for students with LD are significantly longer than the students without LD in all testing conditions and students with LD take longer than their peers without LD to answer implicit question versus literal questions. Also, the group differences for the number of questions correct was usually larger for the implicit question than for the literal questions.

Linder [57] 1989

Adaptation The tests were given under two conditions: Timed and Untimed.
Subjects A total of 100 college age students with and without LD participated: 47 females, 53 males.
Dependent Variable Two tests were used: Scholastic Aptitude Test (SAT), Nelson-Denny Reading Test (1981 edition).
Findings Results indicated that there was no significant difference in the SAT-General Test scores between ability groups. However, testing condition had a significant effect on test performance for both ability groups on the Nelson-Denny Reading Test.

Lord [59] 1956

Adaptation Two level tests (non speeded) and three highly speeded tests were administered.
Subjects Participants were 649 students entering the U.S. Naval Academy.
Dependent Variable Tests of verbal ability, spatial ability, and arithmetic reasoning with 7 tests in each area. School grades served as a criterion measure.
Findings Three factors of the dependent measures were confirmed. A number speed factor and perceptual-speed factor was found and they were separate from a verbal speed and spatial speed factor, which also were found. No arithmetic reasoning speed factor was found. All factors were highly intercorrelated. Grades formed a verbal-academic grade factor.

Mollenkopf [67] 1960

Adaptation Student s took the test first in a highly speeded condition and then were given enough time to finish the test, using a different colored pencil.
Subjects Two groups of high school students participated.
Dependent Variable The number of correct items on two types of measures were used: Verbal analogies, Arithmetic reasoning.
Findings High correlations were found between speeded and nonspeeded tests, although they were considerably higher with verbal tests than with arithmetic tests.

Montani [68] 1995

Adaptation The tests were given under two conditions: Timed and Untimed.
Subjects The students (young children) were divided into several groups: Students with difficulties in mathematics, but not reading (Low Math); Students with difficulties in both reading and mathematics (Delayed); Students with difficulties in reading but not mathematics (Low Reading); Students with no academic difficulties (Control)
Dependent Variable A group of story problems and number-fact problems was used.
Findings Results indicated that the low math group performed worse than the Control group in the timed condition but not in the untimed condition. The delayed group performed worse than the control group overall. Although the low reading group performed worse than the control group overall, the low reading group did not differ significantly from the control group in either the timed or untimed conditions.

Munger & Lloyd [69] 1991

Adaptation The tests were given under two conditions: Timed and Untimed.
Subjects A total of 222 fifth grade students participated in this study, 6 with physical disabilities, 94 with learning disabilities, and 112 without disabilities (125 boys and 97 girls).
Dependent Variable Each student took parallel forms G and H of either the Language Usage and Expression test or the Mathematics Concepts test of the Iowa Tests of Basic Skills.
Findings The results of the study provide no evidence of a difference in test speededness for the group with disabilities and the group without disabilities nor evidence that the groups are differentially affected when the amount of speededness was reduced.

Murray [70] 1987

Adaptation The tests were given under two conditions: Timed, Untimed.
The test was further split into two types of presentation: Two-dimensional, Three-dimensional.

Subjects Thirty students without LD and 30 students with LD (ages 12-14) participated in the study. The students with LD were further divided into two groups: Students with low achievement in both mathematics concepts and mathematics computation (17 students); Students with average scores in both areas (13 students); The JM Spatial Battery, which consists of seven visual-spatial tests, was the dependent measure.

Dependent Variable When time was not a factor in spatial testing, boys without LD and boys with LD with average mathematics achievement performed better on tests of visualization and two-dimensional tests than did boys with LD with low scores in mathematics achievement. There were no significant "between group" differences for these scores under timed conditions. There were also no significant differences among the groups on tests of orientation or on three-dimensional tests under timed or untimed conditions.

Findings When time was not a factor in spatial testing, boys without LD and boys with LD with average mathematics achievement performed better on tests of visualization and two-dimensional tests than did boys with LD with low scores in mathematics achievement. There were no significant "between group" differences for these scores under timed conditions. There were also no significant differences among the groups on tests of orientation or on three-dimensional tests under timed or untimed conditions.

Myers [71] 1952

Adaptation Three forms were administered in different groupings of items to determine if practice on the test affected speed of completion.

Subjects Six hundred midshipmen at the U.S. Naval Academy.

Dependent Variable On a figure classification test, nineteen scores were summarized on attempted and correct items.

Findings The most valid test is one which is moderately speeded and can be completed by 70% of the candidates. On the speeded test, ability and rate of answering form two orthogonal factors.

Ofiesh [72] 1997

Adaptation The tests were given under two conditions: Timed, Untimed

Subjects A total of 60 college students participated in the study: 30 students with LD, 30 students without LD.

Dependent Variable The Nelson Denny Reading Test was used in both the timed and untimed condition.

Findings Results showed that students with LD performed significantly lower on processing speed tests than students without LD. When compared to the students without LD, the students with LD showed a greater benefit from the extended time condition.

Perlman, Borger, Collins, Elenbogen, & Wood [76] 1996

Adaptation The Iowa Test of Basic Skills was administered in two ways: According to the publisher's allotted time of 40 minutes and using an extended time of 2 hours and 30 minutes.

Subjects Participants included 85 fourth (n=28) and eighth graders (n=57) attending 19 schools. All students had IEPs recommending extended time.

Dependent Variable The primary dependent variable was grade equivalent scores on the reading comprehension subtest of the Iowa Test of Basic Skills.

Findings All students in the fourth grade took the test within the publisher's recommended time (34 minutes versus the allowed 40 minutes). Eighth graders took substantially more time (55 minutes instead of 40 minutes). Students achieved higher scores when they took the test with extended time. Gender distributions were similar for both timed and untimed versions of the test.

Powers & Fowles [79] 1996

Adaptation Subjects wrote two essays, one with a 40-minute time limit and one with a 60-minute time limit. Half of the examinees wrote the 40-minute essay and half wrote the 60-minute essay first.

Subjects Study participants were 304 paid volunteers recruited from the pool of examinees who took the GRE General Test between January and May of 1994. Ethnic minority and nonnative examinees were oversampled, and in order to ensure sufficient heterogeneity with respect to writing ability, letters of invitation made a special plea to those students who did not consider themselves strong writers.

Dependent Variable Each participant wrote two full-length essays. The examinee had been sent one topic before the test and encountered the other topic for the first time at the test administration. The essays were scored holistically on a 6 point scale by two trained readers. Questionnaire data were collected regarding perceptions of adequacy of time limits, an estimate of how quickly subjects were able to write, and a judgment of how well the subjects had performed on the writing tasks administered.Subjects also submitted a course-related sample of writing that they had completed as an undergraduate assignment. Several weeks after the administration, subjects completed a 12-item inventory of writing accomplishments on which they indicated which of several writing accomplishments they had made.

Findings Essays written under the 60-minute time limit received moderately higher scores, on average, than did essays written under the 40-minute time limit.

Weaver [108] 1993

Adaptation The test was given in a timed condition, extra time was provided for the students who requested it at the end of the test. In the untimed condition, students were notified that there was no time limit prior to taking the test.

Subjects Eighty-eight college students participated in the study: 39 students with LD, 49 students without LD.

Dependent Variable The Nelson-Denny Reading Test (NDRT) was administered to the students.

Findings Results showed that students with LD obtained significantly lower reading scores on the NDRT than did students without LD. Students with LD derived greater benefit from both extended time and untimed test condition on the Vocabulary and Comprehension subtests of the NDRT than students without LD.

Ziomek & Andrews [115] 1996

Adaptation The test was administered under an extended time condition (up to triple the allotted time as compared to timed condition).

Subjects Over 611,000 student records from 1,006 participating institutions were searched resulting in a total of 2,959 special -tested students matched with valid college GPAs, predicted GPAs, and complete ACT test results Three groups of diagnosed disabilities had a sufficient number of students to warrant further analyses: Attention Deficit Disorder (480); Dyslexia (526); and Learning Disabled (1,258).

Dependent Variable The American College Test (ACT)

Findings The correlation of predicted with actual college GPAs was largest for the attention deficit group regardless of the combination of test package and extended time guideline (r=.45). The correlation between predicted and actual college GPAs is lowest for students diagnosed as learning disabled who were administered the cassette tape under the three hour per test timing guideline (r=.27). The average error of prediction was negative for all but one of the conditions analyzed -- students diagnosed as dyslexic who were administered the cassette version with up to three hours to complete each test had as mean prediction error of .06. Students diagnosed as attention deficit had the largest "relative" over-prediction bias.

Also see four studies in Presentation:Examiner Familiarity [8,21,82,94] and one study in Assistive Devices [33].

Setting: Separate Location and Auditory Stimulation Summary

Setting generally refers to the physical location in which the test is administered, although other environmental conditions may be considered as part of the setting.

Analysis of Literature by Subjects and Test

Three studies have been done on setting accommodations, one of which is summarized here. The other two are located at the end of the section on examiner familiarity. In one of the studies, positive effects were noted for children with ADHD from the introduction of background music while no such improvement occurred for students without disabilities (Abikoff, Courtney, Szeibel, & Koplewicz, 1996). In the study by Derr-Minneci (1990), the oral reading performance of students was higher when they read in their reading group or tested at their teacher's desk over that obtained when tested in an office. Finally, Stoneman and Gibson (1978) found young children with various developmental disabilities improved in their motor imitations when evaluated in a small testing room over that attained when tested in their own classroom (see summary in tables for examiner familiarity).

Analysis of Research Quality and Summary

With such limited research conducted with such diverse setting conditions and outcome measures, it is difficult to generalize to the full range of setting conditions possible during testing. The positive results, however, suggest that this area of research should be more fully explored. Generally, the study by Abikoff, Courtney, Szeibel, and Koplewicz (1996) has been thoughtfully executed and the results are very credible. Other supporting studies in the examiner familiarity section also appear to be sound experimental investigations.

Annotated References of Investigations on Setting:

Separate Location and Auditory Stimulation Summary

Abikoff, Courtney, Szeibel, & Koplewicz [1] 1996

Adaptation Each student was tested over two days with testing during the second day under three auditory stimulation conditions: 10 minutes of music, 10 minutes of background speech, 10 minutes of silence.

Subjects Participants included 40 boys, 20 with ADHD and 20 without (average age of 9-9). Six of the boys with ADHD were receiving Ritalin. Many had a concurrent diagnosis: 4 Conduct Disorder, 9 Oppositional Defiant Disorder (ODD),7 Specific Developmental Disorder (SDD), 4 both ODD and SDD. Nine of the children with ADHD were Caucasian, 8 African American, and 3 Hispanic. In the group of students without disabilities, 14 were Caucasian, 4 were African American, and 2 were Asian.

Dependent Variable

Day One: WRAT-R Arithmetic subtest, WISC-R Vocabulary subtest, Arithmetic Screening Test (AST) was used to determine the child's math skills level

Day Two: Three different 60-problem tests at the child's level were administered. Three scores were generated for each subject: Number of math examples attempted, Number of correct answers, Accuracy (the number of examples answered divided by the number attempted).

Findings The arithmetic performance of the children with ADHD benefited from music when the music condition was presented first. The children without disabilities performed similarly under the three auditory conditions.

See two studies listed under Examiner Familiarity [21, 91]

Presentation and Response - Computer Presentation

Because computers can be used in many different ways that involve any number of different accommodations, this area can be considered as a package treatment that also involves other accommodations. In summarizing this research, computer-assisted testing (CAT) is not addressed as defined by Wise and Plake (1989). CAT "is an assessment process whereby the test is constructed as a test-taker is responding to the item. Selection of items is from a very large, statistically cohesive (unidimensional) item pool based on the test taker's responses to all previous items" (Mills & Stocking, 1996, p. 288). Responses are scored and new items selected to ensure a sufficient reliable judgment about the level of proficiency and skill. Correct responses lead to more difficult items and incorrect responses lead to easier items, thus maximizing the items around the student's true skill level, thus leading to a more reliable estimate of performance. In some of the comparisons between CAT and conventional tests, significant improvements have been reported for college age students (Legg & Buhr, 1992). These findings have been reported with or without the use of the algorithm controlling item presentation sequence (Vispoel, Rocklin, & Wang, 1994), with the possibility of item review (Stone & Lunz, 1994), or the

option to skip or defer answering items (Lunz & Bergstrom, 1994). In all of this research, CAT is reportedly very efficient (in time).

In contrast, the focus on computer-based testing (CBT) "generally refers to using the computer to administer a conventional (i.e., paper-pencil) test" (Wise & Plake, 1989, p. 5). CBT can used to enhance access to tests for students with disabilities because changes are made in the manner in which items are displayed, sequenced, or presented (sequenced or paced).

In most CBT, items are presented singly and individually. Early work by Curtis and Kropp (1961) established that displaying a different number of items on a screen influences students' performance, in part because of the information that items share. They reported significantly higher scores when 1 to 3 items were presented than with the conventional paper and pencil administration (in which many items are presented). Likewise, Hoffman and Lundberg (1976) reported displays of items on a screen has a differential effect on performance with items requiring matching but not with items using multiple choice or true-false formats.

Analysis of Literature by Subjects and Test

Fifteen studies have been published on the use of computer presentations as a test accommodation. Two early studies using the test of Written Spelling, 1976 edition, reported very different results for elementary age students. Whereas Hasselbring and Crossland (1982) found less time needed to take the test with a computer for students with learning disabilities, Varnhagen and Gerber (1984) reported longer time for both regular students and students with learning disabilities. In this later study, more errors were made with the computerized version. Lee, Moreno, and Sympson (1986) found that a computerized test with 30 items was more difficult (for 21 of the items) over a paper-pencil version, with military recruits; however, no persons with disabilities were involved in the study. This same finding of computerized tests being more difficult than paper pencil versions was reported by Watkins and Kush (1988) for 33 elementary grade students with learning disabilities who completed capitalization items. Nevertheless, students were more positive about the computerized test. In a study of reading comprehension, Keene and Davey (1987) reported equivalent performance for both a computerized passage and a printed passage for a group of high school students with

learning disabilities. Another finding in which these two forms have been equivalent – computerized versus paper and pencil – was reported by Miller (1990). In this study, the Peabody Picture Vocabulary test (PPVT) was administered to nearly 100 students, with and without cerebral palsy. Horton and Lovitt (1994) found the computerized administration of a comprehension test to reveal mixed results for middle-and high school students, some of them with learning disabilities. Performance was not the same for factual versus interpretive questions when comparing a paper-pencil and computer version of the test. Finally, for Swain (1997), no interaction was found between students with and without disabilities taking two math tests with and without a computer. Scores for students administered the test with the traditional format, however, were higher than scores obtained under the computerized format.

Burk (1998) has conducted the most recent study with computers used to present the test. She included three major types of accommodation in the presentation of the test: an audio read aloud (for only 1 group of students), large print, and increased spacing. She used several different tests used for high stakes decisions (e.g., the GED, the Maryland Functional Test, and two others) and gave them via the computer primarily to students with disabilities (reflecting two subgroups of students with developmental delays and autism). Her results reflected considerable improvement over that attained when the test was given in the standard paper and pencil format.

Analysis of Research Quality and Summary

The research on using computers in test administration has been extensive over time, with different students in terms of age and disability and using varying methodologies. Because of the rapidly changing nature of computers and the lack of large studies with extensive sampling of students, much of this research is best interpreted with caution. Furthermore, because the use of computers in testing often includes several confounding variables at once (i.e., individual testing, control of item management, etc.), multiple conclusions may be made. Often, the same findings (positive or negative) of using a computerized version of a test appear with such different student groups, tests, or methods, that knowledge really fails to accumulate. Probably the major consistent conclusion to direct future research is the use of computer assisted testing, which is beginning to develop a positive empirical basis. To the degree that students can establish

their own optimal environment when taking the test with a computer, this tool is likely to become a critically important accommodation device. For many students with physical disabilities, computer-based testing may represent the only viable accommodation that gives them access to tests without modifying the meaning of the measure. Yet, almost no research has been reported on this target group of students.

Annotated References of Investigations on Presentation and Response:

Computer Presentation

Burk [14] 1998

Adaptation Three different accommodations were presented via the computer: (a) large print, (b) increased spacing, and (c) audio delivery of problems (for only one group of 12 students out of the total of 182 students from nine different groups.

Subjects A total of 182 students were tested across 8 sites involving primarily students with learning disabilities (n=111), 17 students with developmental disabilities, 4 students with autism, and 50 students without disabilities.

Dependent Variable Four different types of tests were used: The Test of General Educational Development (GED) certification, the Maryland Functional Math Test, Test Ready Materials from Curriculum Associates, the Adult Placement Inventory, and a test for transition used by ARC.

Findings Significant improvements occurred in performance for all six groups of students with learning disabilities and no improvements for students with developmental disabilities or in general education. "In the conditions without sound where print was normal size (comparable to printed tests), extra spacing significantly increased the test scores (by an average of 10.82 points)" (p.11)...[For the one group where the sound worked] "students scored significantly better on the computer; those who had added sound had an average 15 points gain; those without sound had an average 6.25 point gain over paper-based tests. Where paper scores were marginal or just below passing, use of the computer brought scores into the passing column" (p.12).

Curtis & Kropp [18] 1961

Adaptation Test items were projected on a screen one at a time and three items (from least to most difficult) at a time. Both of these conditions were compared to a control condition with students taking the test in test booklets and answer sheets.

Subjects Included a ninth grade class of white students (n=29).

Dependent Variable Several dependent variables were analyzed: Frequency of response, Guessing, Factor responses. The School Ability Test was administered under both normal and experimental conditions. Several tests were administered under the normal conditions: Iowa Test of Educational Development, Form X3S3, Tests 3 to 7; Iowa Silent Reading Test; Clerical Speed and Accuracy of the Differential Aptitude Test; Gordon Personal Profile and Inventory; SRA Primary Mental Abilities, ages 11-17; Thurstone Temperament Schedule.

Findings Both experimental conditions yielded higher means than the control condition. A high correlation existed between all types of administration. Subjects reported no difficulty in responding during the experimental conditions but indicated a preference for three item presentations at a time.

Hasselbring & Crossland [39] 1982

Adaptation Two conditions were used to administer the test:

- Group 1, with 14 subjects, was given the Test of Written Spelling (TWS) in its original form;
- Group 2, with 14 subjects, was given the computerized version of the TWS.

Subjects Participants included 28 students with LD, ranging in age from 9-9 to 14-6. All students were enrolled in a summer school language arts program.

Dependent Variable Two variables were measured:

- Examiner scoring accuracy
- Time required for:
  - -Test directions and administration
  - -Scoring Data Summary
  - -Total examiner time

Findings In all four comparisons, the computerized version required significantly less time than did the written version. Approximately 10 minutes of teacher time per pupil were saved using the computerized assessment, with a net saving of more than two hours for the 14 students involved in this study.

Hoffman & Lundberg [45] 1976

Adaptation Stations were used in a large classroom in which a small box was placed with five numbered buttons. To indicate a response to a question or item, the student simply pushed the appropriate numbered button. The computer then recorded the last response made by the student. The test forms (A and B) were matched item for item in terms of item format, subject matter, and judged item difficulty. Two conditions were used: Using the response system with the box and using a conventional administrative condition. Items were presented visually on 35-mm slides projected on a large screen. Items were read verbally as well as projected. Students were told that they would have approximately 1 minute for each multiple-choice question and 40 seconds for each matching or true-false question. Conventional administrative conditions took place in classroom facilities with a proctor.

Subjects Consisted of 136 Year II pharmacy students enrolled in a lecture -laboratory course in pathology.

Dependent Variable An 80-item, one-hour objective test

Findings The two administrative modes resulted in equivalent scores and test-taking behavior (in terms of number and pattern of changed answers) for the true-false and multiple-choice item formats. However, the sequential, paced mode resulted in significantly different scores and test-taking behavior for the matching item formats. The reduction seen in the mean score of the matching items under the sequential, paced administration was not due to extraneous variables associated with the matching items, but to the manner of presentation of the items.

Horton & Lovitt [48] 1994

Adaptation Two modes of test administration were used: Pencil-and-Paper, Computer.

Subjects Participants included 72 middle and high school students, 38 males and 34 females. Thirteen of the subjects had learning disabilities, 16 were remedial students, and 43 were normally achieving. Of the 29 students with learning disabilities or identified as remedial, 19 were Caucasian, 8 were Asian, and 2 were Hispanic. All were classified as middle class. Three teachers with at least 8 years of experience also participated in the study.

Dependent Variable The measure consisted of nine multiple choice tests, each with 15 questions, 12 factual and 3 interpretive. All questions had four possible answer choices.

Findings Six findings were reported:

- 7% of the students scored substantially higher on the computerized group reading inventory, and 4% performed substantially higher with the pencil-and-paper method. Only three students with learning disabilities displayed a substantial difference, all favoring the pencil-and-paper method on interpretive questions.
- Overall on interpretive questions, the students with learning disabilities scored slightly higher with pencil-and-paper, and the normally achieving students scored marginally higher on the computer assessments.
- On factual questions, the results marginally favored the computer assessments for both the students with learning disabilities and their normally achieving peers. The students with learning disabilities in middle school social studies, however, scored markedly better with pencil-and-paper. In high school social studies, both groups scored marginally higher on the computer assessments.
- Results indicated that students with learning disabilities generally comprehend information as well when questions are presented on computer as when presented from a textbook.
- 70% of the students favored learning information from a computer rather than from a textbook.
- Each teacher preferred using the computer to evaluate their students' levels of independent interaction with the textbook.

Keene & Davey [52] 1987

Adaptation Students read two lengthy expository passages about animals either from print (n=26) or from a computer screen (n=25). While reading, students were encouraged to use one of six reading strategies to help them comprehend the material.

Subjects Participants included 36 male and 15 female students with learning disabilities in grades 9-12.

Dependent Variable Several measures were administered: 14 post-passage questions, each with 4 possible options; A strategy checklist (6 items); An attitude checklist (3 items on usefulness, enjoyment, repetition). Task completion time was also measured.

Findings Students performed equally well with either the computer or the printed page. Students reported looking back on the passage more frequently with the computer screen. Students spent the same amount of time in both conditions. Students wanted to repeat the task more in the computer condition than in the printed page condition.

Lee, Moreno, & Sympson [55] 1986

Adaptation Two modes of test administration were implemented: Paper-and-Pencil; Computer.

Subjects There were 654 male military recruits between the ages of 18 and 25. A total of 334 participated in the paper-and-pencil mode and 320 participated in the computer mode.

Dependent Variable Two measures were used: Arithmetic Reasoning Subtest of the Armed Services, Vocational Aptitude Battery (ASVAB-AR), Experimental Arithmetic Reasoning Test (EXP-AR)

Findings Scores obtained on the paper-and-pencil test were higher than those obtained on the computer test. Of the 30 items, 21 were more difficult in the Computer Mode, while 3 were more difficult in the Paper-and-Pencil Mode. The remaining 6 items were of approximately equivalent difficulty.

Legg & Buhr [56] 1992

Adaptation Two forms of the test were given: Computerized adaptive test (CAT) and 'Conventional' test.

Subjects Participants included 628 community college and university examinees, 57.8% of them female. Mean age was 22.3 years. A total of 63.5% of the examinees were White, 9.7% were Black, 20.7% were Hispanic, and 3.5% were Asian. Most of the examinees had some experience with a computer with 30.5% indicating 'frequent' use, 45.6% indicating 'occasional' use, 20.2% reporting using a computer 'once or twice,' and 3.7% designating that they had 'never' used a computer.

Dependent Variable Two measures were used:

- Mathematics, reading, and writing subtests of the College Level Academic Skills Test (CLAST; State of Florida, Dept. of Ed., 1989)
- A 19-item questionnaire covering the ease in following test-taking procedures, facility in using the computer, constraints of the CAT tests, machine factors, readability, anxiety level, and preference

Findings There were three findings:

- CAT reading scores were about 16 points higher than those for the conventional test.
- Pass-fail decision consistency was very high for both mathematics and writing. Decision consistency was lower for reading.
- Attitudes toward computerized testing were very favorable.

Lunz & Bergstrom [61] 1994

Adaptation Modifications were made to the computerized adaptive testing methodology to make it more similar to traditional paper-and-pencil tests: difficulty of the first item, targeted level of test difficulty, minimum test length, opportunity to control the test.

The opportunity to control the test involved four conditions:

- Skip condition: Examinees were allowed to choose the items they answered
- Review condition: Students were required to answer all items when they were presented but were allowed to review and change item responses after they completed the test
- Defer condition: Students could defer answering items until the end of the test
- None condition: Students had no control over the test and were not allowed to skip, defer, or return to items previously presented

Subjects Subjects were 645 students in medical technology programs.

Dependent Variable A certification examination of 109 items pulled from a 726 item bank was administered to students.

Findings There were no significant differences due to difficulty of the first item, targeted test difficulty, or minimum test length. There was, however, a significant main effect for opportunity to control the test. Students in the "skip" format performed significantly better than students who had no control over their adaptive test. As control over the test decreased from "skip" to "review" to "defer" to "none", the mean ability estimate decreased slightly. The opportunity to skip items was used by 64% of the students who had that option. Students who skipped items earned a lower mean ability estimate than those who chose not to skip items. The patterns of skipping were inconsistent, and students seemed to skip items randomly across difficulty and content. The opportunity to review items and alter responses was used by 61% of the students in that format, and the opportunity to defer items was used by 45% of the students who had that option with students tending to defer the more difficult items.

Miller [65] 1990

Adaptation The tests were given under two conditions: Standard mode of response, Computerized mode of response.

Subjects A total of 96 students participated: 48 students without disabilities and 48 students with cerebral palsy.

Dependent Variable The Peabody Picture Vocabulary Test (PPVT) was given to the students.

Findings The standard mode of response and the mode using a computer were equivalent for the two groups.

Stone & Lunz [90] 1994

Adaptation Computer adapted testing (CAT) was analyzed to determine the effect of item review and alteration, a procedure usually not allowed. Examinees were divided into 3 groups: (a) 1 Standard Error of Measurement (SEM) below passing, (b) Within 1 SEM of passing, and (c) 1 SEM above passing. Examinees also were divided into those who passed and those who failed.

Subjects Participants included 208 examinees taking one test and 168 examinees taking another test, both of which were part of certification for the American Society of Clinical Pathologists.

Dependent Variable The dependent variables were the number of correct items on a certification examination and the decision of passing or failing.

Findings On average, only 3 items were changed on the 50-item test and 4.5 on the 85-item test; the changes went both ways from correct to incorrect and visa versa. Little change occurred in the relative test efficiency after review of items. The confidence of making the same decision after review remained the same for most examinees, particularly in the middle ability group and certainly for both outer groups (well below and well above). In summary, before and after review, estimates of performance were highly correlated.

Swain [93] 1997

Adaptation The tests were presented in a computer format and a paper and pencil format.

Subjects One-hundred fourteen third grade students participated in the study. A portion of the students had disabilities in mathematics, and a portion had no disability.

Dependent Variable The KeyMath-R and the CAMT were administered to the students.

Findings The results revealed no statistically significant interaction between ability group and mode of assessment between the two mathematics tests of similar content. Second, there was statistical significance in the method of assessment used, as evidenced by scores obtained on both formats of the mathematics test than on the computer-administered format of the test. The ability level was a statistically significant factor on both formats of the mathematics test. Subjects who were categorized as normally achieving in mathematics scored higher on all subtests of both tests than subjects who were categorized as mathematically disabled. Also, no mathematical concepts consistently distinguished between normally achieving subjects in mathematics and those who were mathematically disabled.

Varnhagen & Gerber [104] 1984

Adaptation Each student was tested under both a normal written and microcomputer test administrations.

Subjects A total of 27 students participated:

- Eighteen students came from one regular (RG) third grade class (average age of 9-3) with 33% of the students described as non-English speaking with low writing skills.
- Nine students were classified with learning disabilities (LD) and attended a self-contained special education classroom (average age of 11-5).

Dependent Variable The Test of Written Spelling (1976) was administered, a standardized dictation test consisting of predictable (35 words) and unpredictable (25 words) was administered. Students were scored on: Testing Time, Typing/Writing Times, Number of correctly spelled words, and Student attitude.

Findings Students in both the RG and LD groups took longer to respond and made more errors on the computerized test version than on the conventional handwritten version, regardless of the order in which they were tested. All 27 students stated that they would prefer to take future spelling tests on a computer.

Vispoel, Rocklin, & Wang [106] 1994

Adaptation Three formats of computer testing were used:

- Fixed items with students answering the same set of items.
- Computer adapted with item presentations determined by student performance and an algorithm based on passing in relation to item difficulty.
- Self adapted with students deciding on the difficulty of items to be presented.

Subjects Participants included 121 undergraduates from the University of Iowa taking an introductory psychology course. There were 61% women; 88% were White.

Dependent Variable Individual difference variables (test anxiety, academic self concept, and computer usage and anxiety) were monitored to relate them to test performance on three measures: Computerized-adaptive tests (CAT), Self-adapted tests (SAT), Fixed-item tests (FIT).

Findings The CAT was found to be the most efficient, then the SAT, and finally the FIT (which required almost 2 items to every 1 with the CAT). No differences were found between the CAT and SAT. Significant main effects were found for test anxiety and self-concept, which were related to ability; higher estimates were associated with lower anxiety and higher verbal self-concept. The interactions between the other three individual difference variables and administration conditions were not significant.

Watkins & Kush [107] 1988

Adaptation Two types of tests were given:

Computer Test: Mastery criterion was set at 85%, Non-mastery level at 40%, A proficiency ratio (proportion correct) was calculated by the computer and statistically compared, via the sequential probability ratio test, to the pre-specified master and non-master criteria.

Conventional Test: 170 sentence item, Randomly sequenced by objective level, Presented 20 per page in 14 point upper and lower case type.

Subjects Participants included 33 learning disabled students (23 male, 10 female): Average grade placement was 4.5; Average full scale WISC-R IQ was 91; Ethnic representation included 29 White, 3 Hispanic, and 1 Black; 23 were in resource programs.

Dependent Variable The Capitalization Machine software was used to assess the capitalization domain with 17 discrete objective levels. Each of the 17 skill levels contained a pool of 10 sentence items tapping that particular capitalization skill. Student performance on each of the 17 capitalization objectives was characterized as mastery, review, or non-mastery (i.e., >84%, 41%-84%, and <41%, respectively). Scores were summed across all 17 objectives to produce two total capitalization test scores: one for the computerized version and one for the conventional version.

Findings The computerized test resulted in a mean of 25.4. The conventional test version had a mean of 27.7. This difference between test means was significant, with the correlation between scores on computerized and conventional tests equal to .81. Computerized and paper-and-pencil test versions did not significantly differ in their assignment of student to instructional interventions. The computerized test (mean=4.6) was perceived in a more favorable light than the conventional paper-and-pencil test (mean=2.7).

See one Assistive Devices/Supports study [40]

Presentation: Examiner Familiarity

Most large-scale tests rely on administration procedures that are standardized to avoid differences that might arise from examiners. Typically, this standardization is achieved through the use of (a) general guidelines and (b) explicit scripted test directions so no variations are introduced. Such standardization, however, is insufficient in controlling the background relationships between the student being tested and the examiner.

The common research design paradigm is to compare a student's performance with a familiar and then with an unfamiliar person. Often the gender, race, and role of the individual are systematically compared, under varying conditions, using different test performance as the dependent variable. In this research measures of language tend to be heavily used, often involving individually administered tests.

Analysis of Literature by Subjects and Test

This area of research was addressed frequently for a decade in the mid-1970s, with most of the studies conducted by Doug and Lynn Fuchs. Other than two earlier studies by others, a total of 10 studies have been reported by the Fuchs (Fuchs, Dailey, &Fuchs,

1982; Fuchs, Featherstone, Garwick, & Fuchs, 1984; Fuchs, Featherstone, Garwick, & Fuchs, 1981; Fuchs, Featherstone, Garwick, & Fuchs, 1984; Fuchs & Fuchs, 1989; Fuchs, Fuchs, Dailey, & Power, 1989; Fuchs, Fuchs, Daily & Power, 1985; Fuchs, Fuchs, Garwick, & Featherstone, 1983; Fuchs, Fuchs & Power, 1987).

In the two early studies, by Stoneman and Gibson (1978) as well as Olswang and Carpenter (1978), performance was higher with the child's mother present than in the presence of an unfamiliar (same sex) person. In the former study, motor imitation behavior increased for children with Down's syndrome, hydrocephalus, cerebral palsy, and delayed development, while in the latter study, the number of utterances increased for children with language impairments.

The Fuchs' team has conducted most of this research and has consistently reported significant effects when performance is compared in the presence of familiar examiners versus unfamiliar examiners. In this research, the familiar examiner typically is the classroom teacher and the students tend to be young (preschool age), with moderate to severe disabilities. The measures tend to be based on language instruments or behavior samples, with many different types of performance assessed, including imitative behavior, intelligible word production, picture descriptions, labeling, and other gestural or outer directed behaviors. The dependent measures range from counts of simple production responses to scores of syntactic and semantic complexity.

The latest study completed in this area was by Derr-Minneci (1990) who tested elementary students oral reading performance with their classroom teacher versus the school psychologist. Students read significantly more fluently if tested by the teacher.

Analysis of Research Quality and Summary

This research is generally of high quality with the findings unconfounded by many other variables. Typically, the primary independent variable is clearly defined, well isolated and students participate in both conditions (subjects are crossed with treatments). This research is important in the area of accommodations because of the very real probability that many other accommodations are likely to require unfamiliar examiners. For example, if students need extra time, a different room, or any number of alternate schedules to complete the test, it is likely they will receive them through another person, possibly with someone they do not know (an instructional assistant or other classroom

teacher). To the degree that their performance is differentially influenced (positively by the primary accommodation and negatively by the incidental accommodation of unfamiliar examiners), the net effect may not be improved performance. Unfortunately, with most of the findings coming from research on young children, the generalization of the findings to other populations (K-12 students) may be limited.

Annotated References of Investigations on Presentation: Examiner Familiarity

Derr-Minneci [21] 1990

Adaptation The administration of the test was altered in several ways: who administered the test (teacher vs. school psychologist), location of the test (reading group vs. teacher desk vs. office outside the classroom), and duration of the test (timed vs. untimed).

Subjects Participants included 100 third and fourth grade regular education students: 35 students reading below average grade level, 31 students reading at their average grade level, 34 students reading above their average grade level.

Dependent Variable A curriculum-based assessment based on Cones (1981, 1987) elaboration of a methodology for validating behavioral assessment procedures was used. The assessment measured correct words per minute (CWPM) and percentage of errors.

Findings Students read more CWPM when assessed by their own teacher, in their reading group, compared to the teacher desk. Students read more CWPM at the teacher desk compared to the office setting. Timed students read more CWPM than the untimed students. Furthermore, students committed more errors when assessed in an office setting, compared to the teacher desk. Also, the students had more errors when at the teacher's desk as compared to being assessed in the reading group. The location, duration, and tester effects mentioned above were similar across reading levels.

Fuchs, Dailey, & Fuchs [25] 1982

Adaptation Subjects were assessed twice during a period of two weeks–once by a classroom teacher and once by a stranger–within a crossover design. All examiners were female, certified, and had several years' experience.

Subjects Subjects consisted of 34 preschool children, 21 boys and 13 girls, with moderate to profound disabilities in speech or language functioning. The mean age was 4-9. All subjects tested within the normal range on IQ tests.

Dependent Variable Subjects described two pictures from Tester's Teaching Picture Series (1966). Each description was rated as accurate or inaccurate with respect to the content of the stimulus picture and was scored in terms of the total number of intelligible words employed. An 18-category scale was used to measure the semantic/syntactic complexity of the subjects' descriptions.

Findings Subjects demonstrated richer descriptive language, as well as greater fluency, under the familiar examiner condition. Subjects used a greater number of non-repetitive, intelligible words to describe drawings when interacting with familiar rather than unfamiliar testers. The children's total semantic/syntactic complexity score and their complexity score on accurate statements also were greater in the familiar condition than in the unfamiliar examiner condition. Also, children employed a greater number of qualitatively different semantic/ syntactic categories with the familiar examiner.

Fuchs, Featherstone, Garwick, & Fuchs [26] 1981

Adaptation Students were tested by both their classroom teachers (familiar examiners) and by unfamiliar examiners. All examiners were female, certified, and had several years of experience. The administration of the Action Pictures task was also modified – the "No Instruct" condition examiners gave the children adequate time to complete their response while the "Instruct" condition examiners allotted a constant amount of response time.

Subjects Subjects were 79 preschool children with moderate to profound speech and/or language disabilities. Children with mental retardation as well as speech or language disabilities were not included.

Dependent Variable

- The Sounds-in-Words subtest of the Test of Articulation (Goldman & Fristoe, 1972). Neither the subjects' imitative or spontaneous performance were scored since previous findings failed to indicate differential responses to familiar and unfamiliar examiners.
- The Action Pictures (AP) task, during which the students had to describe two ambiguous pictures, was given. Subjects' responses were evaluated in terms of the total number of intelligible words and syllables employed to the describe the illustrations.
- A Symbolic Mediation Test (SMT), a test incorporating three levels of complexity of symbolic mediation (low, medium, and high), was given. At each level of complexity, three items required a verbal response and three items required a gestural response. One point was given for each correct response with 18 as the maximum number of points awarded for the total score.
- The Schenectady Kindergarten Rating Scale (Conrad & Tobiessen, 1967), a 17-item instrument that examines classroom behavior, was given. Each item was rated along a 3 to 7 point scale.

Findings The subjects did not perform differentially when tested by familiar and unfamiliar examiners on the AP task. Student performance on the SMT, however, was significantly better with the familiar examiner than with the unfamiliar examiner. On the AP task, the students' syllabic productions were significantly greater in the 'Instruct' than in the 'No Instruct' condition. Students also employed a greater number of words in the "Instruct" than in the "No Instruct" condition.

Fuchs, Featherstone, Garwick, & Fuchs [27] 1984

Adaptation Students took tests given by both familiar and unfamiliar examiners. All examiners were female, certified, and had several years of experience.

Subjects Participants included 79 (55 males, 24 females) speech- and/or language-impaired preschoolers (average age was 62.35 months). All subjects had been enrolled in one of three special education programs for at least 6 weeks prior to the study. Their performance on standardized language and/or articulation measures ranged from 1.5 to 3 standard deviations below the mean, although they were of at least average intelligence. A total of 68 children were Caucasian, and 5, 4, and 2 were Native American, Black, and Asian American, respectively.

Dependent Variable The Symbolic Mediation Test (SMT), a test incorporating three levels of complexity of symbolic mediation (low, medium, and high), was administered. At each level of complexity, three items required a verbal response and three items required a gesture response. One point was given for each correct response with 18 as the maximum number of points awarded for the total score.

Findings Participants performed significantly better on the SMT when tested by familiar examiners. Their differential functioning did not depend on the task's level of complexity.

Fuchs & Fuchs [28] 1989

Adaptation Subjects were tested under familiar and unfamiliar and unfamiliar examiner conditions.
Subjects This meta-analysis involved 14 studies that included 989 subjects.
Dependent Variable Three types of dependent variables were used: Intelligence tests (7 studies), Speech/language tests (5 studies), educational achievements tests (2 studies)
Findings Caucasian students performed similarly in familiar and unfamiliar examiner conditions. African American and Hispanic students, however, scored significantly and dramatically higher with familiar examiners.

Fuchs, Fuchs, Dailey, & Power [29] 1985

Adaptation Subjects were assessed twice during three weeks, once by a familiar tester and once by an unfamiliar tester. Students were assessed by either two inexperienced testers or two experienced testers. Examiners were female graduate students in either an early childhood education program (inexperienced) or a program for speech clinicians (experienced).
Subjects In this study, 22 Caucasian preschool students (17 male, 5 female) with moderate to profound speech and/or language disabilities were tested. The mean age was 58.3 months. All subjects performed within the normal range on IQ tests.
Dependent Variable Preschool Language Scale (PLS; Zimmerman, Steiner, & Pond, 1979), a comprehensive language test that assesses auditory comprehension and verbal expression; Measures of Examiner Characteristics;
• Role Category Questionnaire (RCQ). Three scores were obtained on the RCQ: (a) number of different constructs constituting the description, (b) degree to which these constructs were interrelated hierarchically, and (c) number of positive, neutral, and negative statements.
• The Attitude Toward Disabled Persons Scale (ATDP; Yuker, Block, & Young, 1966)\Scoring the PLS and the ATDP entailed the summing of the testers' written responses to each instrument's set of items.
Findings Preschoolers with disabilities performed more strongly when tested by personally familiar than personally unfamiliar examiners regardless of the testers' experience. There was no difference between experienced and inexperienced testers' cognitive complexity or in their attitude toward people with disabilities. Also, both examiner groups described people with disabilities relatively simplistically and negatively.

Fuchs, Fuchs, Garwick, & Featherstone [30] 1983

Adaptation Subjects were assessed twice during a period of two weeks–once by a classroom teacher and once by a stranger–within a crossover design. All examiners were female, certified, and had several years of experience.
Subjects Participants included 34 preschool children, 21 boys and 13 girls, with moderate to profound disabilities in speech or language functioning. The mean age was 4-9. All children performed within the normal range on IQ tests.
Dependent Variable Subjects completed two tasks from the Sounds-in-Words subtest of the Test of Articulation (Goldman & Fristoe, 1972): a labeling response task and an imitative response task. The subjects' production of initial-, medial-, and final-position phonemes was analyzed. The third task required subjects to describe two pictures from Tester's Teaching Picture Series (1966).
Findings Subjects employed a greater number of intelligible words on the description task when tested by familiar than by unfamiliar examiners. Subjects did not perform differentially on the labeling or imitative tasks.

Fuchs, Fuchs, & Power [32] 1987

Adaptation Students were assessed twice during a period of 3 weeks, once by a familiar and once by an unfamiliar tester, within a crossover design. Students were assessed by either two graduate or two undergraduate female students enrolled in degree programs in communication disorders.

Subjects Participants included 16 handicapped children of low socioeconomic status (8 with LD and 8 with MR). In both the LD and MR groups, there were five boys and three girls.

Dependent Variable The Clinical Evaluation of Language Functions (CELF) was used to measure receptive and expressive language skills in the areas phonology, syntax, semantics, memory, and word-finding/ retrieval. Outer directedness, as operationalized in terms of the frequency and duration of subjects' glancing behavior, was also measured.

Findings Students with LD performed significantly and dramatically better with familiar, rather than unfamiliar, examiners. Students with MR scored similarly in the two examiner conditions. Students with MR glanced more often and for a longer duration than students with LD. Students with LD regarded familiar examiners more frequently and longer than unfamiliar examiners, whereas children with MR did not exhibit such a difference.

Olswang & Carpenter [74] 1978

Adaptation Two language samples were obtained under two conditions:

- Language elicited by the child's mother
- Language elicited by an unfamiliar, female clinician

The two samples were collected within 1 week, and the order of the collection conditions was counterbalanced.

Subjects For this study, nine children, five males and four females, with language impairments were tested (ages 3 to 6).

Dependent Variable Language samples were obtained during a 25 minute period by the mother playing with her child as she normally did at home or by the clinician by using parallel-play techniques (following the child's lead rather than directing him/her). Only spontaneous utterances were analyzed. Language was analyzed in numerous ways:

- Total number of analyzable utterances
- Vocabulary type-token ratio, the ratio of the number of different words (types) to the total number of words (tokens) in a given sample
- Mean length of utterance, computed by dividing the total number of morphemes by the total number of utterances for each language sample
- Percentage of one morpheme utterances
- Percentage of two morpheme utterances
- Percentage of three or more morpheme utterances
- Proportion of grammatical morphemes per utterance, computed by dividing the number of grammatical morphemes used by the number of utterances in each language sample
- Percentage of occurrence of 13 different semantic categories
- Type-token ratio for each of the 13 semantic categories, the ratio of the number of different utterances in a given semantic relation to the total number of different utterances expressing that relationship

Findings The subjects produced a greater number of utterances when elicited by the mother than by the unfamiliar clinician. The quality of language used with both adults, however, was the same.

Stoneman & Gibson [91] 1978

Adaptation Tasks were presented by the child's mother (familiar examiner) or a female graduate student in special education and/or psychology (unfamiliar examiner). Samples were obtained in both the child's classroom and a small testing room. Each student received all experimental manipulations in one of four predetermined sequences, counterbalanced for possible order effects.

Subjects Participants included eight children, three boys and five girls, with disabilities (average age was 22.5 months). Four had Down's syndrome, one had hydrocephalus, one had cerebral palsy, and the other two had less common syndromes associated with delayed development.

Dependent Variable A seven-item assessment instrument, with four items involving motor imitation skills and three items involving fine motor-manipulative skills was administered. Each child was allowed up to three trials per item, but one correct response trial was sufficient for an item to be scored as correct.

Findings Subjects scored significantly higher on the assessment instrument when administered by their mothers than when administered by unfamiliar examiners. The children also answered more items correctly when evaluated in a small testing room than they did when assessed in their own classroom.

Presentation and Response - Multiple Changes

Often it is difficult to divide test changes as either a presentation or response because one implies the other. In this research, a number of different changes are considered, often studied as a package with several implemented at the same time. Few of the investigations in this section of the paper are confined to one unique accommodation; rather, they are listed multiply across several accommodations.

Analysis of Literature by Subjects and Test

The most frequently studied accommodation is large type and Braille, with 9 studies reported to date. The accommodations were not always on populations with visual impairments. Three large print studies have been conducted with elementary students taking a minimum competency test (Beattie, Grise, & Algozzine, 1983; Grise, Beattie, & Algozzine, 1982; Hidi & Hildyard, 1983). Performance improved with large print, although many other accommodations also were made available, therefore making it difficult to arrive at firm conclusions. Three studies have been reported with college age students taking a college admissions test (Bennett, Rock, & Jirele, 1987, Bennett, Rock, & Kaplan, 1987; and Rock, Bennett, & Jirele, 1988). In these studies, comparability of tasks and items was being investigated (and supported), with certain cautions noted when tests were enlarged.

Studies have been done by three other researchers on a different age group of students. For Coleman (1990), the use of large print presented readers with some difficulty in their writing assessments. While Perez (1980) found improvements in

performance with large print for secondary level students with learning disabilities (LD), Mick (1989) found just the opposite for high school students with LD and educable mental retardation (EMH).

Finally, presentation and response changes have been studied with college age students. Accommodations included extended time, large type, Braille, audiocassettes, readers, assistance in filling in bubble sheets, signing of directions, and use of assistive devices (slate and stylus or magnifying glass). Under standard conditions, performance on the test predicted freshman grade point average (GPA) equally well for students with and without disabilities (r=.59). This relationship was lower for other students with visual or physical impairments or learning disabilities. In a study done by Bennett, Rock, and Kaplan (1987), two instances of differential difficulty were found when students took the GRE in Braille. In contrast, Coleman (1990) reported positive results for 7 students who took a writing test with Braille.

In an analysis of 2,959 special tested students with matching data on valid college GPAs, predicted GPAs, and ACT test results, Ziomek and Andrews (1996) reported on the effect of several accommodations. Three groups were studied, including students with attention deficit disorder, learning disabilities, and dyslexia. Several accommodations were studied: cassettes, regular print, and extended time with two types: double and triple the time on various English, reading, math, and science tests. Generally, little changes in the predictions occurred when students completed the test with accommodations, regardless of disability type. There was a tendency for a slight over-prediction bias. The correlation between predicted and actual GPA was the highest for students with attention deficits and lowest for students with learning disabilities. When analyzing the test package with and without cassettes and extended time, the correlation of errors of prediction and predicted GPAs for students with dyslexia (r=.18) and learning disabilities (r=.17) varied with different test packages.

The ETS studies using the Scholastic Aptitude Test (SAT) or the Graduate Record Examination (GRE) General Test have been summarized by Willingham, Ragosta, Bennett, Braun, Pock, and Powers (1988) for the same targeted groups of students used in the ACT research. The accommodations made by ETS have included alternative test formats (modifying the presentation by using Braille or audio presentations, assistive

devices, and separate locations). They investigated score comparability using five specific indicators (reliability, factor structure, differential item functioning, prediction of performance, and admissions decisions) and task comparability (test content, testing accommodations, and test timing). In general, they found that between the standard and nonstandard administrations, there was...

- comparable reliability, though with some sections of the SAT, the correlations were not as highly correlated for students with disabilities as they were for students without disabilities (Bennett, Rock, & Jirele, 1987; Bennett, Rock, & Kaplan, 1985, 1987).
- similar factor structures with a better fit using a four factor structure than a three factor structure, although the analytical factor did not function as a single factor for students with visual impairments taking a large print version of the test (Rock, Bennett, & Kaplan, 1987).
- similar item difficulties for students with and without disabilities, except for the Braille version of the mathematical test, which had a few more difficult items (Bennett, Rock, & Kaplan, 1985, 1987).
- noncomparable predictions of academic performance (with the nonstandard test scores less valid and test scores substantially underpredicting college grades for students with hearing impairments and overpredicting college performance for students with physical impairments) (Braun, Ragosta, & Kaplan, 1986).
- comparable admissions decisions with minimal effect from flagging, though for students with hearing impairments were more likely to be admitted while students with learning disabilities and visual and physical impairments were less likely to be admitted to smaller institutions (Benderson, 1988).

In an analysis of test content, Willingham (1988) found that students with disabilities performed better on the verbal than on the math sections and although they perceived the test to be harder, they performed comparable to nondisabled peers. Accommodations offered by ETS include Braille, cassette, alternate recording systems, separate test locations, and extra time. However, students with disabilities completed the entire test more often than those without disabilities and college performance was overpredicted when extended time was allowed.

In the end, these researchers recommend that those using any test results " (a) use multiple criteria to predict academic performance of disabled students, (b) give less weight to traditional predictors and more consideration to students' background and nonscholastic achievement, (c) avoid score composites, (d) avoid the erroneous belief that nonstandard scores are systematically either inflated or deflated, and (e) where feasible and appropriate, report scores in the same manner as those obtained from standard administrations" (ETS, 1990, Executive Summary Report). This research, however, is limited to testing with GRE, ACT, and SAT, all of which represent a limited arena for students with disabilities. The proportions of those with disabilities who participate in such tests is very small (proportionately) and may not be representative of the larger group of such individuals (within any disability group); these studies also focus on tests for college-bound individuals and/or young adults.

Six studies have been done in which students had the test read to them. In two of these published reports, the same study reported by two different authors resulted in different conclusions. Koretz (1997) reported on oral reading (along with rephrasing, cueing, and dictation) of math and science tests for 4th and 8th grade students taking the Kentucky Essential Skills Test (as part of KIRIS). He concluded that the test was biased given that students with moderate cognitive and learning disabilities who received the accommodation scored near the mean of students without disabilities and who did not receive the accommodation. In contrast, Trimble (1998) reported that only 4 significant differences appeared from the 104 comparisons that were made comparing students' performance with and without the accommodations. In this research, the reading aloud accommodation was part of a package in which other accommodations also were used (dictation, rephrasing, and cueing) and statistical estimates only were available for documenting its unique affect. In a study by Tindal, Heath, Hollenbeck, Almond, and Harniss (1998), fourth grade students with learning disabilities improved significantly on their math performance when the test was read aloud to them. In fact, they reported a significant interaction, in which no such performance improvement was reported for students without disabilities who received the same accommodation. Fuchs, Fuchs, Eaton, Hamlett and Karns (in press) also reported differentially significant improvement for students with learning disorders over students without learning disorders when adults

read to elementary students a test of math problem solving with extended reading demands; this same accommodation, however, was not effective on a traditional achievement tests of math computation or math concepts and applications which had more modest reading requirements. For Harker and Feldt (1993), high school students' performance was higher on several English and content tests when the passages were read to them; however, students with disabilities were specifically excluded from this sample. Finally, Westin (1998) reported positive differential results when fourth grade students had a math test read to them.

The use of paced item presentation, reductions in items per "page," or video presentations have been studied by four researchers. Curtis and Kropp (1961) conducted a study with general education (9th graders). They found significantly higher performance when items were projected in a paced manner on a large screen (either one or three at a time), relative to taking the test with a traditional booklet and answer sheet (see computer presentation tables for summary). In another computer study, Hoffman and Lundberg (1976) reported decreased performance with sequential pacing for projected items that required matching with the college students. Pacing also was part of the test accommodation reported by Helwig, Tedesco, Heath, Tindal, and Almond (1998), as well as by Tindal, Glasgow, Helwig, Hollenbeck, and Heath (1998). In both of these latter studies, students were paced in their administration of a math test by a video 'read aloud' of the math problems and options. In the former study, students were given the standard test booklet (with multiple problems per page), while in the latter study, students had only 1 problem per page. Although Helwig et. al. found significant effects only for some problem types with the middle school students in the study, Tindal et. al. found significant effects only for the elementary, and not the middle school.

Four studies have been completed with audio presentations of tests, with the improved performance reported in three studies. One study was part of the research done by ETS on college admissions tests, with differential easiness found for students with learning disabilities taking the test with a cassette administration. Espin and Sindelar (1988) reported middle school students identifying more errors when listening to tapes than when reading the passages, though this finding was not differential between the general education students and those with learning disabilities. With high school students,

Perez (1980) reported positive effects from the use of an audiotaped test administration, though performance was less than that attained with the use of large print. Finally, in a series of single case studies, Tindal, Almond, Heath, and Tedesco (1998) reported varied effects when elementary students with learning disabilities took a math test using audio cassettes in which the items and options were read to them. This is one of the few studies in which a "read aloud" was not found to be effective.

Four studies have been reported with changes in the answer sheet, cues on the form, or reduced distractions. Two of these studies have been reported as part of a multi-component modification (Grise, Beattie, & Algozzine, (1982); Beattie, Grise, & Algozzine, 1983) with positive effects. For Veit and Scruggs (1986), 4th grade students with learning disabilities correctly shaded fewer bubbles than did general education students although they were comparable in the number and percentage correct when checked by hand. Peterson (1998), tested elementary age students with and without disabilities using a statewide multiple-choice reading test. He placed questions in close proximity to the passage section to which they were related. His results, however, were mixed: For only 2 of 5 students was any improvement noted, one from general and one from special education.

Research on rephrasing and cueing has been reported earlier with the two studies from the Kentucky state test data, with opposite conclusions reported by Koretz (1997) and Trimble (1998). In the regression analysis by Koretz, the most significant accommodation of the four researched was dictation, with cueing being only modest in its influence. Otherwise, Jackson, Farley, Zimet, and Gottman (1979) found that providing students with a self-vocalization strategy helped students with behavioral and emotional difficulties perform better on the Porteus Maze and WISC-R (see tables for reinforcement).

Olson and Goldstein (1997) note several key issues in their research on multiple test changes for the National Assessment of Educational Progress (NAEP).

1. Accommodating students with disabilities is likely to increase their participation and representation.
2. The test may be different for this group than those in the general population.

3. Issues of comparability remain with previous administrations to determine trend measures.

Two studies have been completed with the level of syntax reduced on the test. Wheeler and McNutt (1983) reported improvements on a math problem-solving test for low achieving 8th grade students, particularly when the syntax was hard (versus moderately hard or easy). This finding was true even with problems at the students reading or computation skill level. Miller (1998) made several changes in the math problems of a statewide multiple-choice test in an effort to simplify the language. She reported only a form effect, finding neither a significant main effect for students with and without disabilities nor any interaction between the type of test (simplified or standard).

Analysis of Research Quality and Summary

The research on presentation accommodations is complex and requires caution in making any firm interpretations. The most clear and positive finding appears to be in the use of large print or Braille and in the use of read aloud of math problems both of which appear differentially effective. Probably the most problematic issue is the great diversity in subjects, tests, and designs that have been used. While this research may have the longest history, the findings are clearer than the conclusions. In general, it appears that making changes in the way tests are presented had a positive impact on student performance although the results have not always been differential for students with disabilities versus those without disabilities. The reasons why, however, are uncertain. Much of this research is done with post-hoc evaluations of extant databases or using quasi-experimental research designs. The accommodations also have been implemented in packages, making it difficult to ascertain specific changes. Finally, the accommodations also have been implemented with many different kinds of students.

Annotated References of Investigations on Presentation and Response: Multiple Changes

Beattie, Grise, & Algozzine [7] 1983

Adaptation Students took one of three tests: modified, modified large-print, and standard. The tests were modified in the following ways:

- Hierarchical progression of difficulty
- Sentences arranged unjustified
- Bubbles (horizontal ovals) were vertically arranged
- Passages were placed in shaded boxes
- Examples were "set off" from test items
- Arrows in the corners of pages that were part of continuing sections and stop signs replaced them at ending pages.
- A complete modified test was prepared in standard type size (12 point) and a second version of the modified test was produced in large print (18 point).

Subjects A total of 345 students with LD students participated in the study.

Dependent Variable The reading portion of the Florida State Student Assessment Test for Grade 3 was the primary dependent measure.

Findings The results suggested that the competence of students with learning disabilities was enhanced by the use of tests which include the modifications mentioned.

Bennett, Rock, & Jirele [8] 1987

Adaptation Three groups of students took the test in three different formats.

- Visually impaired students taking the regular-type edition;
- Visually impaired students taking the large-type, extended time administration;
- Physically handicapped students taking the regular-type edition in normal administration.

Subjects Participants included 339 students with visual impairments, and 151 students with physical disabilities. Two reference groups were used: The first was a group of 441,654 students taking the GRE from Oct. 1981 to June 1984. The second was a group of 20,499 students taking C-3DGR3 under typical testing conditions from Oct.1981 to April 1982.

Dependent Variable The GRE form C-3DGR3 was administered.

Findings With respect to performance level, the groups of students with visual impairments achieved mean scores that approximated or slightly exceeded those of students without disabilities. Students with physical disabilities scored lower on two of the three test scales. Students with physical disabilities and visual impairments taking timed, national administrations were slightly less likely to complete selected test sections than in the other conditions. The reliability of the General Test was found to be comparable to the reference population for all groups with students with disabilities.

Bennett, Rock, & Kaplan [9] 1987

Adaptation The tests in this study were taken in a variety of formats, including: Large type, Braille, Cassette, Cassette and regular type.

Subjects Four groups of students participated in this study: 437 hearing impaired; 6,285 learning disabled; 576 physically handicapped; 1,585 visually impaired.

Dependent Variable The Scholastic Aptitude Test (SAT) was the primary dependent variable: Form WSA3 and Form WSA5.

Findings Five of the 162 pairs (all associated with the Mathematical scale) showed evidence of differential operation for a group of students with disabilities. Two instances were accounted for by clusters that were differentially difficult for students with visual impairments taking the Braille edition of the SAT, whereas in the remaining three instances, clusters showed evidence of differential easiness for students with hearing impairments taking the regular type exam and for examinees with learning disabilities taking the cassette administration. When the individual item performances associated with these five instances of differential operation were examined, no clear indication was provided that these broad item classes function differentially with handicapped examinees.

Coleman [17] 1990

Adaptation The students took the test under one of three conditions: Braille, Large print, Regular print.

Subjects Twenty-four children participated in the study: 7 Braille readers, 7 large print readers, and 10 regular print readers

Dependent Variable Three assessment devices were used: Transitivity of Length, Written Length Assessment, Functional Length Assessment.

Findings Regular print readers had the least difficulty with the tasks and large print readers had the most difficulty. Vision seemed to account for the differences rather than age.

Espin & Sindelar [24] 1988

Adaptation There were two methods for presenting passages and sentences:

- Listening to taped passages and sentences
- Reading passages and sentences

Subjects Students were from grades 6 through 8 (age 10 to 14-7) and included equal numbers of boys and girls: 30 students with LD, 30 same age students without disabilities (CA), 30 same age students with low reading skills (RDG).

Dependent Variable Number and percent of errors identified as well as number of correct items identified as incorrect were the independent variables.

Findings Students listening to the taped passages and sentences identified more errors than students reading the passages and sentences. CA> LD or RDG (effects size = .33). Students with LD identified fewer errors than students in general education. No differences were found in percent correct among the student groups for passages or sentences.

Grise, Beattie, & Algozzine [35] 1982

Adaptation Students took one of three tests: modified, modified large-print, and standard. The tests were modified in the following ways:

- Order
- Vertical format
- Shape of answer bubbles
- Sentences arranged unjustified
- Passages placed in shaded boxes
- Examples "set off" from the test items
- Arrows in the corners of pages that were part of continuing sections and stop signs replaced the ending pages.
- A complete\modified test was prepared in standard type size (12 point) and a second version of the modified test was produced in large print (16 point)

Subjects A total of 344 fifth grade students with LD participated in this study.

Dependent Variable The reading portion of the Florida State Student Assessment Test for Grade 5 (SSAT-I) was used as the dependent measure.

Findings In general, students with learning disabilities performed quite well on the modified versions of the test. The average overall percentage of items answered correctly was over 80%; the participating students' average performance score on the *skills* measured by the SSAT-I (Grade 5) also was greater than 80%. Performance on the regular-print and large-print versions of the test subsections was similar. The performance of students administered the modified SSAT-I was considered better than that of students with learning disabilities administered the standard version of this minimum competency test. Performance was affected by factors of test construction rather than skills, standards, or content of test items.

Harker & Feldt [37] 1993

Adaptation Two conditions were compared: Silently read the test; Silently read the test as it was read to them. Each student participated in both conditions.

Subjects A total of 177 students were selected from five Iowa schools (grade 9, n=114; grade 10, n=3). Students were specifically sampled to avoid using any students with a learning disability or receiving special education services.

Dependent Variable The Iowa Test of Educational Development was administered, with several subtests scored: social studies, reading, interpretation of literary materials vocabulary, and use of sources of information.

Findings Performance with the taped administration was significantly higher than the "read silently" group (.7 standard score point). This effect was greater for the Interpret Literary Materials test while no difference was found with Vocabulary. For the Social Studies and Use of Sources of Information subtests, the effect of a taped administration interacted with student reading level (it improved performance for low and middle readers). Correlations were high among the subtests across the two types of administration.

Helwig, Tedesco, Heath, Tindal, & Almond [41] 1998

Adaptation Students answered 60 math multiple choice items during 2 testing sessions of 30 questions each. No calculators were allowed on either version.

- One half was presented in a standard test booklet. Students read the questions silently and responded by circling the appropriate answer in the booklet (testing was self-paced);
- The other half was presented via a video monitor with a person reading the test while the students followed along with a test at their desk (testing was reader paced).

Subjects Thirty-three students from 15 sixth grade classrooms were involved in the study. Four of the classes were conducted in resource rooms for students who had been identified as needing assistance in math. These classes ranged in size from 7 to 12 students. The 11 regular education classrooms contained from 18 to 34 students. Students were predominantly White from low to middle socioeconomic backgrounds.

Dependent Variable Students answered 60 math multiple choice items during 2 testing sessions of 30 questions each. The total number of items correct was calculated.

Findings Data analysis was performed on four subgroups: Low Reader/High Math, Medium Reader/High Math, High Reader/High Math, Low Reader/Low Math. For each group the difference was calculated in success rate between the standard and video version of the test on each item. The values were correlated with 8 passage attributes. While the correlations tend to be highest for the Low Reader/High Math group, only one value reached statistical significance. As the number of verbs present in a passage increased, the difference in success rate in favor of the video accommodation tended to increase. For six test items identified as complex, one half of the items showed significant differences between the two test formats. Two of the items favored a video presentation while the remaining item favored a standard presentation.

Koretz [54] 1997

Adaptation The most frequently used accommodations were grouped into four major classes: Dictation, Oral reading, Rephrasing, Cueing. Data were examined on these accommodations singly and in combination along with actual test performance.

Subjects Participants in this study include fourth and eighth grade students in special education participating in the Kentucky statewide assessment program.

Dependent Variable Two major dependent variables were used: the frequency with which an accommodation was used and the performance on the statewide test in math and science when the accommodation was used

Findings In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Mick [64] 1989

Adaptation An unmodified test was compared to a modified test which had: Increased print (14 point), Unjustified lines, and Responses in booklet, not on bubble sheet.

Subjects The study included 85 secondary special education students: 36 with LD (29 boys and 7 girls ages 15-0 to 17-5 and IQs of 75-109); 40 with educable mental handicaps (EMH) (26 boys and 19 girls ages 15-1 to 18-8 and IQs of 42-84).

Dependent Variable The dependent measure was the Virginia Minimum Competency Test: IOX Basic Skill Reading Secondary level: Understanding Safety Warnings, Completing Forms and Applications, Using Common Reference Forms, Determining Main Ideas, Using Documents to take Action

Findings Students with LD and EMH performed significantly better on the unmodified test than on the modified test:

- 17 students with LD passed the unmodified version and 13 passed the modified version
- 2 students with EMH passed the unmodified version and 2 passed the modified version

Miller [66] 1998

Adaptation Use of simplified language in math multiple choice items were changed by changing passive to active voice, replacing unfamiliar with familiar words, shortening long clauses and making conditional clauses into sentences.

Subjects A total of 47 students participated with most of them Caucasian from low socio-economic backgrounds and attending 3 different low math classes offered by 2 teachers. All students were judged to be average in intellectual functioning, with 14 students having IEPs in math.

Dependent Variable A statewide multiple-choice math test sampling 12 different math functions was administered, along with a maze reading test, a math computation skill test, an open-ended math problem-solving test, and a math vocabulary test.

Findings Although no order effect was found, the two versions of the test were different (form A versus form B). In separate analyses of each form, no significant differences were found for the accommodated and standard version of the test. No differences were found between general and special education students and no interaction was found between student classification and accommodation. Moderate correlations were found across the various reading and math tests.

Olson & Goldstein [73] 1997

Adaptation A variety of accommodations were made in the administration including extra testing time, multiple sessions, individual or small group administration, reading the directions, giving answers orally, using special mechanical apparatus, using large print and large face calculators and Braille.

Subjects Students with disabilities were included in the NAEP testing if they had an IEP and the multidisciplinary team thought they could participate in the testing with accommodations.

Dependent Variable National Assessment of Educational Progress Mathematics measures.

Findings The assessment was harder and less discriminating for students with disabilities. Most of the items had lower percent correct statistics and smaller item-total correlations. Higher rates of omissions were apparent for students with disabilities. Students with disabilities responded differently than students in the total sample. The effect of new inclusion criteria isn't as pronounced as that of offering new accommodations. While students with disabilities can be included with accommodations, the effect on trend measurement will be difficult to evaluate.

Perez [75] 1980

Adaptation The test was presented in three formats: Regular print, Large print, Audio support.
Subjects The test was given to 48 secondary-level students with learning disabilities (LD).
Dependent Variable The Florida Statewide Student Assessment Test was used in the study.
Findings Large-print presentation resulted in the highest levels of performance overall. Performance with large print was significantly higher than performance with regular print and significantly higher than performance with audio support. Audio presentation resulted in higher performance than regular-print presentation. The large-print presentation resulted in higher performance levels than the regular-print presentation for five of the eight skills tested, and in higher performance levels than audio support for four of the eight skills tested.

Peterson [77] 1998

Adaptation Questions on a reading multiple-choice statewide test were redistributed so that they were close in proximity to the passage information needed to answer them.
Subjects Items from two statewide tests were sampled, consisting of passages with an average of 450 words and with readability levels that were arranged to be comparable across accommodation phases.
Dependent Variable Five students were tested, 2 were in general education, 1 in Title 1, and 2 students were with learning disabilities and IEPs in reading. All students were tested using a withdrawal-reversal design.
Findings The results were mixed with the accommodation effective for 1 general and 1 special education student; it was equivocal for the student in Title 1; and it may have resulted in lower performance for 1 general and 1 special education student.

Rock, Bennett, & Jirele [82] 1988

Adaptation The tests in this study were taken in a variety of formats, including: Students with visual impairments taking the regular-type edition; Students with visual impairments taking large-type extended-time administration; Students with physical disabilities takes the regular-type edition in a timed, national administration.
Subjects A total of 447 students participated: 339 students with visual impairments; 108 students with physical disabilities. A reference group of 20,499 students took the regular-type edition under typical testing conditions.
Dependent Variable The GRE General Test Form C-3DGR3 was used as the primary measure for both the modified and typical administration.
Findings This study investigated the comparability of General Test Verbal, Quantitative, and Analytical scores for disabled and non-disabled populations. A simple three-factor model based on Verbal, Quantitative, and Analytical scale item parcels was posed. Analytical scale scores did not have the same meaning for these two groups of students with disabilities as they did for students without disabilities. The contribution of the factor scores in determining performance on the test's general factor deviated for the two groups mentioned above. Such differences suggest that the use of composite scores should be avoided. Some indications of lack of fit for the Quantitative scale were detected. Across several solutions, the group of students with visual impairments taking the large-type test exhibited considerably lower interrelations with the Verbal factor than did the remaining groups.

Tachibana [94] 1986

Adaptation The tests were presented in either a visual or auditory format. The students were also given extra time in one of the conditions.

Subjects The subjects were 45 community college students with learning disabilities. The students were tested to distinguish between visual, auditory or no preference in regards to modality strengths.

Dependent Variable The test used was the Reading Comprehension Test (The College Board).

Findings A significant disordinal interaction occurred between modality strength groups (visual and auditory) and the test modalities. Students with no preference performed better on the auditory modality. Students improved scores significantly on both the visual and auditory modalities of the reading test given additional time. Students also performed much better on the first halves of both the visual and auditory test modalities than on the second halves.

Tindal, Almond, Heath, & Tedesco [98] 1998

Adaptation Using a single subjects design with a modified multiple-baseline across subjects, pairs of students had a math test read to them using an audio cassette player. They were tested individually and monitored carefully by an examiner.

Subjects Students were in the fourth grade and attended public school. Most students were from low to middle socioeconomic backgrounds.

Dependent Variable The number of items answered correctly in sets of 5 were counted over successive days during 4 phases: baseline, read aloud, baseline, and read aloud.

Findings Performance was not significantly enhanced with math problems being read and student performance being monitored individually. For a few students and within some problem sets, however, performance improved.

Tindal, Glasgow, Helwig, Hollenbeck, & Heath [99] 1998

Adaptation Students took a 30-item multiple-choice math test in which a videotape was used to present four accommodations: problems and options were read; as options were read, they were color cued; only one problem was presented on each page; students were paced through the entire set with predetermined solution times set for each problem. Students also took a standard 30-item multiple choice math test in which they read the problems silently and paced their completion. Finally, several criterion measures (in reading and math) were administered to correlate with performance on either of the above math tests.

Subjects About 2,000 students (463 in special education) from 10 states participated in this study, with just fewer than 1,000 in grades 4-5 and just more than 1,000 in grades 7-8. Across most sites and states, an equal percentage of males and females participated. Most students were White, with lower percentages of Black students, and even lower percentages of other ethnicities (except in 2 states). A substantial percentage of students were receiving either free or reduced price lunch, though these percentages varied considerably among the states. Virtually 100% of the population spoke as the primary mode of communication.

Dependent Variable The dependent variable was the total number of items correct (total of 60 for both versions). The criterion measures included: Reading maze - Number of words correctly selected from 5 options; Math computation skill - Number of problems solved correctly on a production task; Math vocabulary - Number of math words correctly chosen as synonyms to a math word stem; Open-ended math problem - Holistic rating of answer quality. Also both teachers and students were surveyed about perceptions of skills, capabilities, and task-test features.

Findings Providing students with disabilities a video taped read aloud of a math test is a valid and useful way to provide an accommodation in multiple choice test administration. At the very least, it does not detract from performance and has the potential for improving performance with some students. It is likely that these results are a function of specific demographic characteristics of the students, the context of the region (educational and geographic site), and/or the background of the student. Even when no significant group gains are present, however, it is likely that the intervention worked for some individual students. This study utilized a group design and reported average performance reflecting a nomothetic approach. Another strategy involves analyzing individuals using an ideographic approach. Further analyses therefore are planned to understand the data pattern from both perspectives.

Tindal, Heath, Hollenbeck, Almond, & Harniss [100] 1998

Adaptation Reading and math multiple-choice tests were completed either by bubbling an answer sheet or directly marking the booklet. Math tests were orally read in their entirety by the test administrator, including the general directions (for filling out the forms and taking the test), each specific problem, and all items for multiple choice problems. The reading of the math test was standardized.

Subjects A total of 481 students in fourth grade from 22 classrooms in seven buildings were included as subjects. The average age was 10.3 with a range from 9 to 12 years old. Most students were White, 3.7% were Hispanic, and very few other ethnicities were represented. Most students (97%) reported English as their first language. A total of 78 students were being served in special education (171 different IEP areas).

Dependent Variable The dependent variable was the percent correct on the statewide math and reading test (total correct possible was 30 per format).

Findings General education students performed significantly higher than special education students in reading and in math. For both tests, performance was not higher when students were allowed to mark the booklet directly than when they had to use a separate bubble sheet. Students in special education with IEPs in reading or math performed significantly higher when the math test was read by teachers, rather than when they read the test themselves. In contrast, the performance of the 10 lowest achievement-ranked students in general education revealed no improvements when teachers orally read the math test over that achieved when students silently read the math test.

Trimble [102] 1998

Adaptation The analysis was designed to display the relationship between accommodation use and performance over time (4 years) and across various accommodations (singly or in combination with each other): reader/oral, scribe/dictation, cueing, paraphrasing, interpreter, technological, other

Subjects Participants for this analysis included approximately 4,000 students with disabilities who participated in each of three grades (4, 8, and 11) in the Kentucky statewide assessment system.

Dependent Variable The major dependent variables were: type of accommodation implemented; classification of the student into one of four levels of proficiency in math, social science, and science; level on an equated scale score in math, social science, and science.

Findings Data indicate that students with disabilities are improving at a very rapid rate, particularly in some grades. The gap separating them from general education students is closing. The percentages of students participating in the statewide test increases over time but decreases over grades. The percentages of students receiving accommodations is quite high (62% in grade 11 and 84% in grade 4). The performance of students with disabilities was lower than the performance of students in general education yet the impact on the total grade level outcome was minimal given they represented only 10% of the population. In only 4 of 104 uses of accommodations was the performance of students higher than the performance of the total group (with paraphrasing and dictation). In some cases, performance is lower with the use of accommodations than with no use.

Veit & Scruggs [105] 1986

Adaptation Students were given three subtests of the Comprehension Test of Basic Skills (CTBS) and had to fill in a bubble sheet.

Subjects Students for the study were 101 grade 4 students, ages 119 mos. to 130 mos. with 19 students with learning disabilities (14 boys and 5 girls) and 82 students without any disabilities (47 boys and 35 girls).

Dependent Variable The number of items marked correctly on the bubble sheet and percent outside the bubble on the CTBS.

Findings Students with disabilities scored fewer items correctly than the general education students, but were the same in the number of items inside the bubble. No differences were found in the percent of correct items or percent of items marked outside the bubble.

Weston [110] 1999

Adaptation A mathematics test was read aloud and performance on it was compared to that attained with an alternate form of the mathematics test taken under standard conditions (students read the test by themselves). The test had items reflecting two levels of reading difficulty (with text and without text involving calculation only.

Subjects A total of 121 fourth grade students participated: sixty-five with learning disabilities and fifty-six without disabilities. They participated in both conditions, which were counter-balanced in order of administration.

Dependent Variable The mathematics test served as one of the dependent variables. Also, the Terra Nova Reading test (third grade level) served as a criterion measure for scaling reading skill and ascertaining the covariation of performance enhancements. Ratings and rankings by teachers, as well as interviews with them, also were used to study the effects of the accommodation.

Findings Although significant main effects were found for both groups of students and both forms of the test, an interaction also was found rendering the effect mute: A larger effect from the accommodation was found for students with disabilities. "Much of the effect for learning disabled students occurs at lower reading levels where regular education students are not well represented in the this study" (p. 9). Finally, both types of math problems showed an effect from reading: the calculation as well as the word problems.

Wheeler & McNutt [111] 1983

Adaptation Three tests composed of increasingly more difficult sentence structures were administered: The Easy Syntax Test (EST), The Moderate Syntax Test (MST), The Hard Syntax Test (HST)

Subjects A total of 30 eighth grade students in remedial mathematics classes participated: 19 males, 11 females, ages 13-6 to 15-1. Of this group, 29 students were Caucasian and 1 was Arabian.

Dependent Variable The problems on all tests were selected from a fourth grade mathematics textbook and required addition and subtraction computations with and without regrouping. The tests contained only words or derivatives of those words considered to be at or below a fourth grade reading level according to the Dale-Chall List of 3000 Familiar words.

Findings Syntactic complexity affects low-achieving eighth grade students' abilities to solve mathematical word problems. The study revealed no significant differences between the EST and MST, but did indicate significant differences between both these tests and the HST. Also, syntactic complexity may affect students' abilities to solve mathematical word problems even when the problems are at the students' computational and reading-vocabulary levels.

See two Computer Presentation studies [18, 45], a Reinforcement [49] study, and an Assistive Device/ Support-calculator [33] study. See Ziomek & Andrews [114] 1996 in Timing/Scheduling of Testing

Response: Dictation to a Proctor or Scribe

This area of research is aimed at allowing students without writing skills to take the test, generally because of cognitive difficulties as well as physical problems. Voice capture software would fall into this same type of accommodation. This accommodation generally applies to tests in which the student is required to write an answer; typically such tests are writing samples, though with increased emphasis on problem-solving and extended response items in mathematics and content areas, this accommodation may become more important. Dictation to a proctor or scribe also is sometimes described as an oral response.

Analysis of Literature by Subjects and Test

Of the six studies that have been completed on this type of test change, three have been reviewed previously under presentation changes (Fuchs, Fuchs, Eaton, Hamlett, & Karns, in press; Koretz, 1997; Trimble, 1998). In all three studies, dictation appeared to be an effective accommodation, though it has not been possible to ascertain its separate effects. For Fuchs et al., the test was a math problem-solving test with extended writing demands, administered to elementary students. For Koretz (1997) and Trimble (1998), the Kentucky Essential Skills Test was studied with 4th, 8th, and 11th grade students. Fuchs, et. al. found differentially large benefits for students with learning disorders over effects for nondisabled peers. Using multiple regression to obtain an optimal estimate of each single

accommodation and then comparing predicted performance with the accommodation to that without the accommodation, Koretz reported dictation to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively. Trimble's (1998) conclusions on the same data-set are not as definitive, though two findings appear relevant. First, in 4th grade for the 1995-1996 database, "dictation, and the paraphrasing & dictation & other combination, produced mean scores above the total" (p. 24). Second, dictation was present in all 4 accommodations in which performance of students with its use was higher than the total group of students. In both KIRIS studies, the accommodations were used with students from many disability categories.

In a study conducted much earlier than any of these three investigations, Hidi and Hilyard (1983) reported that dictation was an effective accommodation for improving elementary students' writing production. No students with disabilities, however, participated in this study. A similar finding was reported by MacArthur and Graham (1987), with longer and better compositions produced with dictation by elementary age students with learning disabilities. Similarly, Higgins and Raskind (1995) and Raskind and Higgins (1995) reported positive effects for college age students with learning disabilities who composed or edited essays, respectively, with a speech recognition system. In the former study, speech recognition was compared to dictation to a human transcriber and to a control condition of writing without any assistance. In the latter study, students found more errors on a proofreading task than in either the read aloud or no assistance conditions.

Analysis of Research Quality and Summary

The use of dictation consistently appears to boost performance across the range of skills tested, with content performance tests that measure general achievement as well as with compositions reflecting measures of written expression. The change appears to improve performance with a wide range of students as well, for students with and without disabilities as well as for students attending elementary schools and college. Other than the study by Higgins and Raskind (1995) and Higgins and Raskind (1995), who used a voice recognition system, however, little information is presented in this research on the

selection or training of scribes or the rules used to change speech into text. Without this information, the use of scribes may include a wide range of variance across individuals, both students and scribers. Furthermore, much of the naturalistic research reported on the KIRIS reflects post-hoc evaluations with weak internal validity. These designs do not offer conclusive evidence either in support or in criticism of dictation as an accommodation. These research designs are in contrast to the experimental the studies that used voice recognition systems or the studies by Fuchs et. al. (in press) or Hidi and Hilyard (1983), all of which employed adequate experimental controls. While the MacArthur and Graham (1987) study assigned students to each of the accommodations, so few students were tested that further research is needed on dictation or word processing. As voice recognition software becomes increasingly sophisticated, policy makers will have to struggle with the extent to which such technology "levels the playing field" or provides an unfair advantage.

Annotated References of Investigations on Response: Dictation to a Proctor or Scribe

Hidi & Hildyard [42] 1983

Adaptation Students wrote both an opinion essay and a narrative essay and were randomly assigned to one of two types of response modes for each type of essay: Oral, in which students spoke out their composition (opinion) with no time limit (with individual administration); Written, with students given the opinion essay topic and told to write as much as they could (again, no time limit with group administration).

Subjects Participants in this study included 20 grade 3 students and 23 grade 5 students.

Dependent Variable Essays were scored for: Semantic well-formedness (rating of 1-5), Cohesion (rating of 1-5), and Number of words produced.

Findings For both grade levels, children produced better narrative than opinion essays, although no significant differences existed between oral and written essays. On cohesion, the scores on narrative essays were significantly higher than the scores on opinion essays. Fifth graders produced more than third graders. Also, the oral productions had more words than written productions, and opinion essays contained more words than narrative essays.

Also see two studies in Presentation [54, 102], two studies in Assistive Devices [43, 62], and a study in Assistive Devices/Support [33].

Alternative Response

While most tests require still require bubbling a response to a multiple-choice item, other formats have begun to appear in the testing arena. With the advent of performance assessments has come extended response items and scoring systems that provide credit

for strategies and partial answers. Portfolios, especially for writing, also have become increasingly popular.

Over the last decade, the area of performance assessment has increased in both research and practice, with a plethora of formats and systems being promulgated. Generally, this change has come about because of the distrust in multiple-choice tests (see publications of Fair Test) and the research on curriculum-based measurement with different scoring systems and scales using performance tasks (see Tindal, 1998a, Tindal, 1998b, summarizing this work). Quite likely, with IDEA '97 and its mandate for students with disabilities to fully participate in large-scale testing, this kind of research and development is likely to increase even more. At present, however, few validity studies have been reported in the literature.

Analysis of Literature by Subjects and Test

Six recent studies have been conducted, only two of them published in the literature with two presented at conferences. Dalton, Morocco, Tivnan, and Rawson (1994) and Dalton, Tivnan, Riley, Rawson, and Dias (1995) compared several different kinds of test formats with an electricity unit for elementary age students. They reported positive outcomes with a constructed diagram test for students with learning disabilities and a hands-on performance task for students without learning disabilities. Supovitz and Brennan (1997) also reported positive outcomes in their comparison of standardized tests to portfolios for a sample of 1st grade students, they also found, however, several racial inequities between the two formats for 2nd grade students as well as gender differences also were found. No students with disabilities were included in their sample. In two studies reported by Braden, Elliott, and Kratochwill (1997) and Elliott and Kratochwill (1998a, 1998b), discrepant results appeared for both students with disabilities and those without disabilities, both groups having taken several different content performance tasks. Finally, recent research by Arick, Nave, and Jackson (1997) indicates that performance tasks allowed students with severe learning disabilities and generalized cognitive deficits to better exhibit their skills, with the effects stronger in middle schools than in elementary schools.

Analysis of Research Quality and Summary

The research comparing student performance on traditional versus performance assessments is becoming extensive but is being conducted primarily in general education settings. At times, this research reflects the difficulties inherent in constructing comparable tasks, the problems in getting reliable judgment, and the issues inherent in conducting experimental research. To the degree that research is yet under review for publication, many issues about research methodology are still uncertain. At the very least, more research and development needs to be completed before common definitions begin to appear and critical features are identified.

Annotated References of Investigations on Alternative Response

Arick, Nave, & Jackson [4] 1997

Adaptation Performance on standardized assessments and performance-based assessments was compared.
Subjects The participants were 275 students with IEPs and 296 randomly selected students without IEPs.
Dependent Variable Four measurement tools were used: performance-based assessment data, standardized assessments, school records of student performance, and school surveys of participating teachers
Findings The students with IEPs scored higher on performance assessment tasks than on traditional standardized assessments. In several content areas, performance scores of high school students with IEPs were significantly higher than elementary school students with IEPs, while the performance scores of students without IEPs remained constant. The students with IEPs scored lower than students without IEPs on performance tasks.

Braden, Elliott, & Kratochwill [10] 1997
Elliott & Kratochwill [22] and [23] 1998a, 1998b

Adaptation Performance on three types of tasks was compared: knowledge tasks, on-demand performance tasks, and teacher constructed classroom tasks.
Subjects A total of 600 students participated (from 9 school districts). Subjects included fourth and eighth grade students with and without exceptional needs.
Dependent Variable Students were administered: Open-ended math and science performance assessments; Multiple choice measures in math and science.
Findings Performance instruments varied in unpredictable ways and did not fare well in convergent and divergent validity analyses. Measures using the same format but from different subject areas often correlated higher together than differently formatted tests from the same subject matter. On almost all achievement measures, students with Exceptional Educational Needs (ENN) scored lower than their non-ENN counterparts. One performance assessment showed higher scores for students with ENN over students without ENN.

Dalton, Morocco, Tivnan, & Rawson [19] 1994

Adaptation The students were assessed in two ways:

- A written questionnaire included eight multi-part textual questions that asked students to use writing and drawing to convey their understanding and applications of scientific concepts
- A diagram test also included eight questions. It focused on simple, series, and parallel circuits and conductors/insulators. For seven of the eight questions students were asked to predict whether a bulb would light by checking "yes" or "no." Students were to give a brief explanation of their answer

Subjects A total of 172 fourth grade students participated in this study, including 33 with learning disabilities.

Dependent Variable A written questionnaire and a diagram test were given to assess students' understanding before and after instruction.

Findings For fourth grade students in a hands-on science program on electricity, the effect of test format appears to be a function of both learner status and level of domain knowledge. These results suggest that students with LD and students with low and average academic skills are better able to access and use their knowledge in a constructed diagram format that in an open-ended questionnaire format. In contrast, high achieving students appear to be less sensitive to these format differences, performing comparably on the two types of assessment. In addition, this study suggests that graphics may be more useful than textual questionnaire items in helping students who have less domain specific expertise to access and use their "fragile" knowledge.

Dalton, Tivnan, Riley, Rawson, & Dias [20] 1995

Adaptation A science unit on electricity was tested using several different formats: Paper and pencil questionnaire using writing and drawing; Constructed diagram using figural materials with brief written or drawn explanations; Multiple choice (MC) using visual and verbal formats; Hands-on Performance with 5 individually administered tasks.

Subjects Subjects for the study included 74 fourth grade students, 29 with learning disabilities; students were diverse in terms of ethnicity and language background.

Dependent Variable Each of the measures was scored differently: ratings with National Assessment of Educational Progress (NAEP) criteria and ratings with anchor sheets of 1-4

Findings Results indicated that students with LD scored higher on the constructed diagram than on the MC and questionnaire (comparable scores were obtained for the MC and questionnaire). Students with and without LD performed more strongly on the hands-on tasks than on any of the paper-and-pencil measures. Students with LD scored lower than their peers on all these test formats except the constructed diagram test.

Supovitz & Brennan [92] 1997

Adaptation The study made comparisons of the influence of student gender, race/ethnicity, and socioeconomic factors on scores _obtained on: standardized tests and portfolios

Subjects A total of 5,264 first and second grade students participated in the study.

First Grade	*Second Grade*
• Average age = 7.22	• Average age = 8.28
• White = 23%	• White = 22%
• Black = 60%	• Black = 58%
• Latino = 15%	• Latino = 18%
• Other = 2%	• Other 2%
• Free/reduced lunch = 80%	• Free/reduced lunch = 80%

Dependent Variable The dependent variable was student standardized test scores [the California Achievement Test (CAT-5) for first graders and the Degrees of Reading Power (DRP) for second graders] and language arts portfolio performance (scored using a 9 point scale in reading and writing).

Findings The study provided some evidence that portfolios are more equitable than standardized tests. For first grade students, however, portfolio assessment was not more equitable than standardized tests. In the second grade the opposite was true. There were differences in equity associated with membership in specific racial/ethnic and gender groups. Black students performed better relative to White students on the portfolio than they did on the standardized tests. Latino students performed better than the Black students on the standardized tests, but not different from Black students on the portfolio assessments. The gender gap for the portfolios was larger when compared to the gender gap on standardized tests, with girls performing better than boys by a wider margin on the portfolio than on the standardized test.

Response: Mark Responses in Test Booklet

With this accommodation, students mark the booklet directly rather than shade a bubble on a separate answer sheet. The use of multiple-choice tests and Scantron® technology with separate bubble sheets has enabled broader sampling and easier scoring of large-scale assessments. With this advancement, however, comes the danger that students get lost and begin incorrectly aligning the questions from a test booklet with the same corresponding item on the bubble sheet. The result is an incorrect and invalid score.

Analysis of Literature by Subjects and Test

Four studies have been completed with this change in testing with three of them previously reported under presentation accommodations. These three studies reported earlier by Mick (1989), Veit and Scruggs (1986), and by Tindal, Heath, Hollenbeck, Harniss, and Almond (1998) showed the change to be either negative or not effective for students with learning disabilities. While Mick's study was done on a published achievement test and with secondary students, the Tindal et. al. study was done on a

state test and with elementary students. For Rogers (1983), no effects were found for students with hearing impairments taking a 50-item multiple-choice spelling test. Performance was the same whether responding on a separate answer sheet or in their test booklet.

Analysis of Research Quality and Summary

In all four studies, students participated in both conditions (were crossed with the accommodation), with a sample of students with learning disabilities or hearing impairments included in sufficient numbers to conduct a comparison. In all three studies, this accommodation was neither effective overall nor differentially for students with disabilities. All studies done on this type of test change have been conducted with sufficient internal validity to support the conclusion: Making test booklets is an accommodation that at worst is ineffective and at best effective only for individual students.

Annotated References of Investigations on Response: Mark Responses in Test Booklet

Rogers [83] 1983

Adaptation Two adaptations were made:

- Book Response (Answer blocks were placed to the left of the alternatives. Students indicated their response by shading with a pencil the appropriate answer space).
- Separate Answer Sheet Mode (Answer blocks were replaced by the letters A, B, C and D). The answer sheet contained three answer rows for the examples and two 25-item columns for the test questions. Answer spaces were arranged horizontally with the letters A, B, C and D placed immediately above the corresponding space. (Students indicated their response by shading with a pencil the appropriate answer space).

Subjects A total of 156 students participated in the study. Students met the following criteria: 8 to 16 years of age with average hearing loss of 60dB or greater.

Dependent Variable The dependent measure was a 50-item, multiple choice, self-administered spelling test that comprised a disproportionate random sample of the series *Spelling in the Language Arts*, for grade levels two through seven (excluding three letter words).

Findings In agreement with the findings for hearing students, the reliability of test scores is not significantly altered when hearing impaired and deaf students, 8-10 years of age and older, respond to achievement test items on separate answer sheets rather than directly in their test books. Also, the data gained by means of an answer sheet appear to be valid; test scores were not adversely affected when answer sheets were used.

See two studies Presentation [64, 100]

Response: Work Collaboratively with Other Students

In most testing situations, students work alone. In these test changes, they work in small groups, generally to prepare them to take the test. With the use of cooperative groups in many elementary and secondary classrooms, it is a natural extension to consider testing in the same manner that students have been taught. And given the further development of inclusion programs and many different kinds of peer tutoring systems, test taking in a comparable manner may allow students with disabilities better access to the test itself.

Analysis of Literature by Subjects and Test

Four studies have been done with collaborative grouping of students showing that performance is affected in a way that may question whether the outcome is reflective of the individual or the group. For Webb (1993), active participation in the group was a critical factor and the scores from the two test administration conditions were not comparable. In this study, middle school students took a math open-ended problem test. Saner, McCaffrey, Stecher, Klein, and Bell (1994) reported the same findings of non-comparability, though their study on the California Learning Assessment System failed to describe the population of students. Fuchs, Fuchs, Karns, Hamlett, Katzaroff & Dutka (1998) administered mathematics performance assessments to students in individual or paired formats; students were classified as below-, at-, or above-average on math. Results suggested a poorer relationship between the performance assessments and traditional tests when those assessments were completed cooperatively, rather than individually, for below-grade students. In the only study done in this area that included students with disabilities, Poplun (1996) also reported considerable differences in project and individual scores for fifth graders taking an achievement test.

Analysis of Research Quality and Summary

Results suggest that cooperatively completed assessments may not represent student capacity well, especially for lower-performing students, such as those with disabilities. These effects appear consistent across the methodological dimensions of the studies.

Annotated References of Investigations on Response: Work Collaboratively with Other Students

Fuchs, Fuchs, Karns, Hamlett, Katzaroff, & Dutka [31] 1998

Adaptation Performance assessments were administered individually or in pairs.

Subjects Participants were 131 fourth-grade students (55% male; 28% African-American, 64% White, 7% Asian).

Dependent Variable Mathematics performance assessment, scored on a 6-point rubric along four dimensions: conceptual underpinnings, computational applications, problem-solving strategies, and communication.

Findings Among individually administered measures, correlations with criterion measures were moderate and significant; correlations were stronger for performance assessments individually rather than cooperatively completed; and exploratory analyses suggested that cooperative performance measures were more accurate for above-grade-level students than for below-grade-level students.

Pomplun [78] 1996

Adaptation The test was administered individually and in groups.

S ubjects Participants included 888 fifth grade students with disabilities. In this group, 68% were male and 32% female; 85% were White, 7% were Black, 4% were Hispanic, 2% were Native American, and 1% were Asian.

Dependent Variable Two measures were analyzed: A 30-item objective achievement test (on a scale of 10-50 points) and an 11-item science attitude scale (1-4). Group cooperation was rated by teachers

Findings Correlations between scores for project scores and individual scores were somewhat lower for students with disabilities. Groups containing a student with disabilities scored higher than predicted given the individual project scores, individual group scores, and group cooperation scores.

Saner, McCaffrey, Stecher, Klein, & Bell [87] 1994

Adaptation Students were paired in groups to provide assistance in solving a science problem.

Subjects No information is provided; a note is made that 30-40 scorable responses were available for each assessment.

Dependent Variable The dependent variable was comprised of a hands on science exercise from the California Learning Assessment System: a recycling problem (grade 5) or a search for gold (grade 8). The problem was solved over 3 days and was scored holistically on a 1-5 point scale.

Findings Scores obtained while working in a group cannot be interpreted in the same way as scores obtained when working alone. In grade 5, performance in the group was not correlated with performance from prior individual work; in grade 8, this correlation was high. Final individual performance for lower performing students is in part a function of their own work and that of their partner.

Webb [109] 1993

Adaptation Students were grouped heterogeneously by their teacher (3 to 4 per group) to work collaboratively for one 50-minute class.

Subjects Participants included 53 seventh grade students (55% female and 45% male). Of this group, 66% were Hispanic, 21% were White, 11% were African American, and 2% were Asian American.

Dependent Variable Alternate forms of an open-ended math problem were administered individually and in the collaborative group. Protocols were scored by counting the number of correct steps, providing a score of 0-9 points. Student participation in the groups was analyzed by level of involvement (no assistance, showed difficulty and received assistance, and did not contribute to group discussion)

Findings Student participation in the group was predictive of the scores in the group solution and explained the lower scores for many students on the individual test. Both student skill and group participation correlated highly with individual test performance. Students needing help scored better on the individual test when all students participated in the group. Scores from the group assessment are not valid indicators of individual test performance and participation in the group must be known to explain individual test performance.

Assistive Devices: Word Processors

Computers tend to be used in schools primarily as word processors, providing a range of editing and publishing features to assist in the writing process. This use of computers in language arts programs may force test developers to reconsider how writing tests are administered and forcing curriculum writers to reconsider how writing is taught. Therefore it is not surprising that by far the assistive device studied with the greatest frequency is the use of the word processor in writing tests. While direct measures of writing are quite consistently implemented in large scale testing, the influence of word processors in this process is not well understood. Typically the task involves a composition being generated in response to some type of prompt that generates a specific discourse. However, in using word processors in writing tests, all dimensions of the process need to be considered: the task, the process, and the judgement.

Analysis of Literature by Subjects and Test

The findings from using word processors in writing tests are contradictory. For example, MacArthur and Graham (1987) found that the use of word processors did not help 5th and 6th grade students with learning disabilities achieve higher scoring essays or use words more correctly than they achieved when they hand wrote their essays. In contrast, Vacc (1987) reported positive outcomes for four middle school students

described as mildly mentally retarded finding that their compositions of letters were enhanced by the word processor. Two studies also have reported contradictory findings on the scoring outcomes from typed versus handwritten compositions. Significantly higher ratings were given to college students' typed compositions over those given to handwritten compositions in a study by Arnold, Legas, Obler, Pacheo, Russell, & Umbdenstock (1990). In contrast, handwritten essays had received higher scores than typed essays for college age students (Powers, Fowles, Farnum, & Ramsey, 1994) and for middle school students (Hollenbeck, Tindal, Stieber, & Harniss, 1998). In a final study of scoring systems, Helwig, Stieber, Tindal, Hollenbeck, Heath, and Almond (1998) reported on the unidimensionality of writing for middle school students within mode (either handwritten or word processed) and across traits (six different scores). The type of computers used to write with also has been shown to be an important variable. For Hollenbeck, Tindal, Harniss, and Almond (1998), even AlphaSmart® computers were effective tools for middle school students. Likewise, Hollenbeck, Tindal, Heath, and Almond (1998) reported differences whether the spellchecker was used or whether the composition was typed only on the last of 3 days for middle school students.

The use of computers to provide speech synthesis also has been reported by Higgins and Raskind (1995) and Raskind and Higgins (1995) for college students to compose essays and then proofread them to find errors. This assistive device was very effective in improving both the compositions and the error checking process.

Analysis of Research Quality and Summary

This literature is generally quite recent and requires replication. Because so many variables are present in this line of research, the only solution to establishing an empirical basis is to develop a systematic program of research. The variables that need to be considered include: (a) type of discourse, (b) type of computer and monitor, (c) type of writing process, (d) use of various word processor features (spell and grammar check, editing capacity, speed, etc.), (e) rater training, monitoring, and judgment, and (f) student experience. Although the research done to date is intriguing in identifying variables in need of additional research, few experimental designs have been used. Furthermore, the age range is generally constricted to middle school students and older age high school or college students. Finally, in analyzing this test change, it is critical to keep in mind that

the goal is to compensate for behavioral or cognitive limitations brought on by the manner in which the test is administered or taken and not just to improve performance.

Annotated References of Investigations on Assistive Devices:

Word Processors

Arnold, Legas, Obler, Pacheco, Russell, & Umbdenstock [5] 1990

Adaptation Raters judged the writing quality of both handwritten and word processed essays. The 300 handwritten essays were word processed.

Subjects The participants were students attending a college that had taken a placement exam and final exam consisting of a writing sample.

Dependent Variable A holistic score was given to each paper, ranging from 1 to 6. Raters and students completed surveys about attitudes and judgments.

Findings Handwritten papers were rated lower than word processed papers. Long papers composed by a word processor were rated significantly higher. Raters preferred to read handwritten papers. Students hand wrote because of typing skill deficits or word processed because of editing features.

Helwig, Stieber, Tindal, Hollenbeck, Heath, & Almond [40] 1998

Adaptation Two modes of writing presentation were evaluated: paper-and-Pencil, and computer. Students had taken a statewide writing test by handwriting compositions that were later transcribed into a computer with a word processor and printed. State trained raters evaluated both the handwritten and the typed papers.

Subjects Participants in this study were 117 students from seven eighth grade classrooms. The sample was predominately White with an even split between females and males. Ten students were receiving special education services, all with learning disabilities. Most students had used computers before.

Dependent Variable The compositions were evaluated using a scoring guide that ranged from 1 to 6 in quality. Six traits were scored: Ideas-Content, Organization, Voice, Word Choice, Sentence Fluency, and Conventions

Findings A series of factor analyses were completed, first separately; a single factor was found in each , one for handwritten and one for typed. When analyzed together, two factors were found: one for handwritten and one for typed. Factor analyses were performed on the following groups:

- students who did not use spell-checkers
- students rated average or below in writing proficiency by their teachers
- students rated high by their teachers in writing proficiency
- students who were frequent or regular computer users
- males and females
- those students who hand wrote or word-processed imaginative essays

Higgins & Raskind [43] 1995

Adaptation Participants wrote three essays, one for each of the following conditions:

- Using a speech recognition system
- Dictating the essay to a human transcriber
- Without assistance

Students were allowed to handwrite or word process the 'no assistance' essay but were not allowed to use the spell-checking function.

Subjects Subjects were 29 post-secondary students with learning disabilities enrolled at California State University, Northridge (CSUN). Twenty-three students were Caucasian, 3 were African-American, and 3 were Hispanic. The mean age was 24.9 years and the mean IQ was 97.

Dependent Variable Students wrote essays from one of six possible questions. Essays were holistically scored on a scale of 1 to 6.

Findings Speech recognition assists students with learning disabilities in compensating for their difficulties in written composition. When compared to receiving no assistance, students achieved higher holistic scores using the technology. Speech recognition apparently allowed students to use their more extensively developed oral vocabularies at a level that was statistically significant.

- The single most sensitive predictor of the holistic score was words of seven or more letters.
- The program was much better at making correct guesses for longer words than for short, unisyllabic ones.
- The ratio of unique words to words was negatively correlated with composition length, a powerful predictor of holistic score.

Hollenbeck, Tindal, Harniss, & Almond [46] 1998

Adaptation Students were assigned by classrooms to one of two groups:

- Students used a computer to complete the entire test (from planning to final essay). There was no spellchecker available for use by the students.
- Students used an AlphaSmart® computer to complete the entire test (from planning to final essay). There was no spellchecker available for use by the students.

Subjects Seventy-eight seventh grade students (46 females and 32 males) participated in this study; two were served in special education and 76 were served in general education. Their average age was14.3 years old. Most of the students were European American (n=67), with 2 African Americans, 1 Hispanic, 2 Native Americans, and 5 Asian/Pacific Islanders. Students averaged about 10 absences for the year; the mean grade point average for the year was 2.8.

Dependent Variable The compositions were evaluated using a scoring guide that ranged from 1 to 6 in quality. Six traits were scored: Ideas-Content, Organization, Voice, Word Choice, Sentence Fluency, Conventions

Findings The AlphaSmart® groups' mean scores were significantly higher than the Computer Group's scores for all six traits.

Hollenbeck, Tindal, Stieber, & Harniss [47] 1998

Adaptation Raters judged essays in two forms: Handwritten, Typed (the handwritten essays transcribed by the researchers). Each judge did not rate both forms of the same essay. The judges did not know that the typed essays were originally handwritten.

Subjects Subjects were 80 middle school students (7 students were receiving special education) with an average age of 15.1. Half of the students were male and half female. 94% of the students were Caucasian.

Dependent Variable The writing portion of the Oregon Statewide Assessment was administered. The compositions were evaluated using a scoring guide that ranged from 1 to 6 in quality. Six traits were scored: Ideas-Content, Organization, Voice, Word Choice, Sentence Fluency, Conventions

Findings Analysis showed that the original handwritten compositions were rated significantly higher than the typed composition on three of the six writing traits for the total group. Further, five of the six mean trait scores favored the handwritten essays.

MacArthur & Graham [62] 1987

Adaptation Each student composed three stories, one using each method of text production: Handwriting (HW), Word processing (WP), Dictation. Each story was composed in response to a colored picture.

Subjects Participants included 11 fifth and sixth grade students (six male and five female, six black and five white). Each student had been identified as having a learning disability and attended a special education resource room program for approximately one hour a day.

- Mean age–143.6 months
- Mean test performance on Test of Written Language (TWL)–82.59
- Word processing experience:
  - -Nine subjects wrote on word processors in their regular class
  - -Five used a word processor at home
  - -All but two used a word processor at least once a week for 30 minutes

Dependent Variable Follow-up interview during which students were asked: to pick the stories their friends would like best, which method they preferred and why, and whether HW or WP helped them write better and why.

Language Complexity: Number of words, Average T-unit length, Corrected type-token ratio, Number of different words divided by the square root of 2 times the total words, Proportion of mature words, Proportion of grammatical errors to total words

Mechanical Errors: Spelling, Capitalization, Punctuation

Quality and Story Structure: A holistic evaluation procedure on an 8 point scale; Schematic structure of eight story grammar elements–main character, locale, time, starter event, goal, action, ending, and reaction

Time and Rate Measures: Pre-writing time, composing time, and composing rate

Revisions: Syntactic level of the change, such as word or sentence, and the type of operation, such as addition or deletion; Five levels were analyzed: surface, word, multi-word, T-unit, and multi-T-unit

Findings The results demonstrate that dictation differs considerably from both handwriting and word processing. Dictated stories were significantly longer and of higher quality. They also had fewer grammatical errors. For students with LD, the mechanical and conventional demands of producing text appear to interfere with the fluency and quality of written expression. When these demands are removed via dictation, students with LD compose more fluently and with better results. Dictation was approximately 9 times faster than handwriting and 20 times faster than word processing. In contrast to the observed differences between dictated and written stories, no significant differences between handwriting and word processing were found on any of the product measures. Handwritten and word processed stories did not differ on length, quality, story structure, mechanical or grammatical errors, vocabulary, or average T-unit length. Although word processing was less than half as fast as handwriting, the overall amount of revision was similar for handwriting and word processing, as was the syntactic level of the revisions.

Powers, Fowles, Farnum, & Ramsey [80] 1994

Adaptation Study 1: Students produced at least two essays - one in handwritten form and one on computer. Students selected from a pair of topics (personal experience or general issues) on which to write for 50 minutes. The handwritten essays were then word processed so that they resembled those that were originally produced on the computer.

Study 2: Major modifications in training raters included: emphasis that handwritten and word-processed essays may make different impressions, the influences of perceived length on essay scoring, using both handwritten and word-processed essays in the training, and checking for differences in the standards.

Subjects A sample of 32 writers was drawn from a larger sample of 568 students in college. Most were White (71%), with Black (17%), Asian (7%), and other (5%) minority students included. Students reported a variety of majors.

Dependent Variable All essays were scored independently on a 1 to 6 scale by two trained readers using holistic scoring methods to generate scores. The scoring guide on which readers were trained emphasized such qualities as clarity of expressions, logical organization, effectiveness of style, ability to support ideas, and control of grammar and mechanics.

Findings Essay readers gave higher scores to handwritten essays than to word-processed essays. This result was found when examinees' essays were originally handwritten and then converted and re-scored as word-processed essays, and also when original word-processed essays were converted and re-scored as handwritten essays. The results of the Study 2 revealed a smaller effect of the mode in which essays were scored. This effect was the same regardless of the direction of conversion.

Raskind & Higgins [81] 1995

Adaptation Subjects were given the choice of writing by hand or using a word processing program (without spell-checking). Subjects returned for a second session to proofread and locate errors in their essays under three conditions:

- Using a speech synthesis/screen review system (SS) that enabled subjects to select text on the computer screen and hear the words spoken as they were simultaneously highlighted. It was possible to review the text by word, line, sentence, or paragraph. Students could modify the rate of speech, volume, pitch, and the colors of the background and highlighted text for maximum contrast and readability.
- Having the text read aloud by a human reader (RA)
- Having no assistance (NA)–proofreading the hard copy independently
- No time constraints were placed on the subjects in any of these conditions

Subjects The study included 33 students with learning disabilities (19 male and 14 female) at California State University, Northridge (CSUN). Students were 19 to 37 years old with a mean age of 24.9 years. Subjects were predominantly Caucasian and middle class with 25 identified as Caucasian, 4 as Hispanic, 3 as African-American, and 1 as Asian-American.

Dependent Variable Subjects wrote an essay of three to five typewritten pages on a topic of their choice or from a list of six topics. Nine categories were used to score the essays: Capitalization, Punctuation, Spelling, Usage, Grammar, Mechanical, Grammar-Global, Typographical, Content/Organization, Style. The total number of errors found by each subject was divided by the number of errors found by the raters. This resulted in the percentage of total errors found by each subject for each condition.

Findings Results indicated that under the SS condition, subjects found significantly more of the total errors (35.5%) than in either the RA (32%) or the NA (25%) conditions. The difference between the RA and the NA condition also was significant. The use of a speech synthesis system also outperformed the other two proofreading conditions in seven out of nine categories of written language errors–four of them at a statistically significant level.

Tindal, Hollenbeck, Heath, & Almond [101] 1998

Adaptation Students took a statewide writing test that required them to create a composition over 3 days using either: Paper-and-Pencil or Computer.

Students who took the test with a computer were assigned to various groups: (a) composing on the computer for all 3 days, (b) composing on the computer only the last day, and (c) composing with a spellchecker available.

Subjects A population of 164 seventh grade students participated in this study; 44 were served in special education and 120 were in general education. Of the students for whom the record data was available, their average age was 13.3 years old, 54 were females, and 58 were males. Most of the students were European American (n=89), with 1 African American, and 4 Hispanic.

Dependent Variable The compositions were evaluated using a scoring guide that ranged from 1 to 6 in quality. Six traits were scored: Ideas-Content, Organization, Voice, Word Choice, Sentence Fluency, Conventions

Findings No significant differences were found for all six traits between the handwritten and computer generated essays. No significant difference in form of computer use was found for four traits: ideas and content, organization, voice, and word choice. However, for both sentence fluency and conventions, the mean Computer Group score was significantly lower than the Computer-Last-Day with SpellCheck Group, which also outperformed the Computer-Last Day group on all traits except ideas and voice.

Vacc [103] 1987

Adaptation Students completed letters by hand and with the word processing program Wordstar on an Osborne microcomputer. Students had 45 minutes per letter writing session.

Subjects Participants included one white and three black male students in eighth grade, certified as mildly mentally handicapped (MMH, with WISC-R full scale at 72). All participants had been enrolled in a special education program for at least 2 years and had completed a one-semester course in typing. The mean age was 15-1.

Dependent Variable Several dependent variables were measured: Time to complete the letter, Number of revisions mad, Letter length, Words per minute (the number of words in the letter divided by the time needed to complete that letter), Quality of each letter using a holistic guide, Effect sizes for each dependent variable.

Findings Significant differences existed between the two treatment modes for all dependent variables except quality. All four subjects spent significantly more time completing letters, wrote substantially longer letters, and undertook a greater amount of revising when composing letters on the microcomputer word processor. All subjects wrote more words per minute when completing handwritten letters and no differences were found in the quality of letter between the writing mode. Time to complete a letter was positively correlated with the total number of revisions made per letter and the length of letter.

Assistive Devices: Calculators

Calculators have been examined within math tests when the assessment involves more than simple calculation. As the standards for math have incorporated a process approach (see the National Council of Teachers of Math Standards, 1989), less emphasis has been placed on computation skills and fluency. Rather, conceptual underpinnings have been emphasized. Administration of mathematics problems with the aid of

calculators make sense if the goal is to measure a deeper understanding without the restrictions of rote calculation. However, if math problems are developed to test such rote calculation, then calculators would undermine the meaning of the test and should not be allowed. This area of research has focused, therefore, on which types of problems should allowing calculators for use with specific groups of students.

Analysis of Literature by Subjects and Test

Of the four studies completed, generally consistent findings have been reported in the outcomes. For Loyd (1991), calculator use was determined by the type of problem: On occasion, it was helpful; at other times, it was not helpful for high school students. This same finding was reported by Cohen and Kim (1992) and Bridgeman, Harvey, and Braswell (1995) with college students. In a similar way, on a test of math problem solving, Fuchs et. al. (in press) reported a marginally significant effect from the use of calculators: Although both performance for students with and without learning disorders decreased with calculators, the decrease was smaller for students without learning disabilities. On tests of conventional math concepts and applications content, again scores decreased for students with and without disabilities; this time, however, the decrease was smaller for students without disabilities.

Analysis of Research Quality and Summary

The studies, although few in number and some of limited sample size, provide a common-sense conclusion that targets the effect of this test change as a function of problem type. Clearly, some math problems require rote calculation and allowing a calculators to be used would invalidate any judgments of student proficiency. At other times, the calculator is rendered useless or harmful, again as a function of problem type. Most of the research has been well-designed and executed, although the studies have been confined to older students. These studies collectively, suggest caution regarding the use of calculators for students with LD.

Annotated References of Investigations on Assistive Devices: Calculators

Bridgeman, Harvey, & Braswell [13] 1995

Adaptation Half of the sample took the test with the use of a calculator while the other half took the test without the use of a calculator.

Subjects The sample consisted of 11,457 college-bound high school juniors from a total of 257 high schools. In 19 of these schools, 40% or more of the student body consisted of African Americans.

Dependent Variable Four measures were used:

- A 70-item test that included all item types proposed for the mathematics portion (SAT-M) of the new Scholastic Aptitude Test (SAT) (i.e., regular mathematics, quantitative mathematics, and student-produced response)
- A questionnaire to determine the extent of calculator use among the students, what types of calculator the students normally used, and whether they used calculators on school math tests
- Students also went back through the test to mark questions on which they had used a calculator and noted whether they thought the calculator had been very helpful, somewhat helpful, or not helpful for each of those questions. In the no-calculator, students indicated items that they thought would have been easier with a calculator, using the same 'helpfulness' categories as above
- A background information questionnaire

Findings Results indicated:

- The use of calculators resulted in a modest score increase, although effects on individual items ranged from positive through neutral to negative.
- Calculator effects were found on items at all difficulty levels, and calculators were beneficial for students at all ability levels. Prior experience in using calculators in testing situations appeared to be very beneficial.

Cohen & Kim [16] 1992

Adaptation On 2 forms, items were classified into 4 groups: Computation only, Computation possible but answer is misleading, Algorithm needed to solve without computation, Algorithm and computation needed. Students used calculators only on the second half of the form.

Subjects Participants included 1490 students (765 on form 1 and 725 on form 2) enrolled in calculus and precalculus math courses.

Dependent Variable The dependent measure was a 28-item test using operational items from the precalculus sections of a standardized university mathematics placement test.

Findings Calculator effects were detected in 12 items:

- On 2 items computation problems were easier with calculators
- On 2 items, the function key on the calculator made them easier
- On 8 items, use of a calculator impeded performance due to inappropriate calculator use rather than lack of mathematical skill.

Loyd [60] 1991

Adaptation A math test was given with 4 subsets of items: I-Easier with a calculator, II-No great advantage with a calculator, III-Calculator not needed, IV-More difficult with use of calculator

Subjects Participants included 160 high school students ages 13 to 17 (69 boys and 71 girls). The majority of students were White (83%) with Black (10%) and other (7%) ethnic backgrounds present. Most (60%) students were very comfortable using a calculator.

Dependent Variable A 32-item math test was used as the dependent variable.

Findings Three major findings were reported:

Almost 50% of the students didn't use a calculator. Calculator use was predicted by item type:

- For Type I problems, students with calculators performed better than those who did not use them
- For Type II problems, the use of calculators did not result in significantly greater performance
- For Type III problems, use of calculators was not helpful
- For Type IV problems, no difference appeared between students with or without calculators
- More time was needed to use calculators.

See Time/Scheduling [33]

Other: Reinforcement

The basic definition of reinforcement is the introduction of a stimulus following a response with the effect of increasing the likelihood of the response occurring again. Since Skinner defined operant conditioning and the field of applied behavior analysis began, a multitude of studies have been conducted on the effects of reinforcement in shaping and conditioning the occurrence and rate of a range of social, communicative, and academic behaviors. No one seriously doubts the applicability of reinforcement in maintaining behaviors, yet such an analysis has had only a brief and sporadic influence on test taking.

Analysis of Literature by Subjects and Test

Investigations of reinforcement techniques in testing situations have been dominated by two major research teams with several individual studies being conducted. In a program of research aimed at understanding reinforcement for students from differing backgrounds, the following studies have established that estimates of intelligence are dramatically influenced when (a) various types of reinforcement are used (Terrell, Taylor, & Terrell, 1978), (b) the race of the examiner and the type of reinforcement is considered (Terrell, Terrell, & Taylor, 1980), or (c) the reinforcement is either tangible or culturally relevant (Terrell, Terrell, & Taylor, 1981). Another team of researchers have established that race, socio-economic status, and reinforcement interact such that (a) white children

improved in their estimates of intelligence with both immediate and delayed reinforcement (Young, Bradley-Johnson, & Johnson, 192), (b) black and white children from low socioeconomic conditions differed in the effects of immediate and delayed rewards (Bradley-Johnson, Johnson, Shanahan, Rickert, & Tardona, 1984), (c) black children classified as educable mentally retarded improved in their performance on the WISC-R (verbal only) with tokens (Johnson, Bradley-Johnson, McCarthy, & Jamie, 1984), and (d) immediate tangible reinforcement was very effective for young children from low socio-economic conditions (Bradley-Johnson, Graham, & Johnson, 1986).

Other independent researchers have established the positive effects from (a) using immediate (versus delayed or noncontingent) reinforcement to improve performance on the Raven Progressive Matrices (Smeets & Striefel, 1975), (b) the differential effects using reinforcement versus feedback on IQ tests, which was ironically contradictory even for the same subtests (Willis & Shibata, 1978 and Jackson, Farley, Zimet, & Gottman, 1979), and (c) the positive effects of both verbal praise or tokens over neutral verbal statements (Saigh & Payne, 1979).

Finally, Koegel, Koegel, and Smith (1997) have conducted one of the best contemporary studies on the effects of reinforcement in changing test performance. They found that young students diagnosed with autism performed better on standardized language and intelligence tests when specific motivation and attention conditions were addressed during the test administration.

Analysis of Research Quality and Summary

This research has been executed with classical group designs in which students have been either crossed or nested within treatments using either counterbalancing or random assignment, respectively. Sample sizes have been adequate and treatments generally well defined and specifically administered. Finally, a strong focus on students with disabilities has been prominent. In great part, such high quality research is more easily achieved when behavioral phenomena are being studied. A clear conclusion can be reached from this research: Performance on standardized tests, whether considered as aptitude (a.k.a. intelligence) or achievement, is greatly influenced by the contingencies invoked (either directly manipulated or implicitly present) when the test is taken. As with all

reinforcement strategies, however, they are not specific to subgroups of students with disabilities.

Annotated References of Investigations on Other: Reinforcement

Bradley-Johnson, Graham, & Johnson [11] 1986

Adaptation The test was administered under two conditions: Standardized procedure, Token reinforcement immediately following each correct response.

Subjects The study included 20 control subjects (10 first- and second graders, 10 fourth and fifth graders) and 20 experimental group subjects (10 first and second graders, 10 fourth and fifth graders). All subjects were Caucasian and of low socioeconomic status.

Dependent Variable Two tests were administered to each child: Slosson Intelligence Test (1975) to make sure groups had equal IQ ranges, WISC-R Intelligence Test.

Findings The children who received immediate tangible reinforcement scored significantly higher on WISC-R Verbal, Performance, and Full Scale scores than the children who received only the standardized administration procedures.

Bradley-Johnson, Johnson, Shanahan, Rickert, & Tardona [12] 1984

Adaptation The test was administered under three conditions: Standard administration, Standard administration with delayed reinforcement, Standard administration with immediate reinforcement

Subjects *Experiment 1*: Participants included 33 black, second grade children attending an inner city, public elementary school. All children were from families of low socioeconomic status.

Experiment 2: Participants included 33 white, second grade children, all from families of low socioeconomic status.

Dependent Variable Two tests were administered: Slosson Intelligence Test (1963), a 10 minute verbal test, to assess equality of the groups with respect to IQ; WISC-R Intelligence Test except the Mazes subtest.

Findings *Experiment 1:* Results show that black, low-income second graders scored an average of 13 IQ points higher in the immediately reinforced condition than the black children tested under standard conditions or with delayed reinforcement.

Experiment 2: The white, low income, second graders scored eight points higher under the immediate reinforcement condition, but this was not statistically significant and was not higher than the delayed reinforcement group.

Jackson, Farley, Zimet, & Gottman [49] 1979

Adaptation The test was administered under five conditions: Reinforcing attention; Reward for success; Self-vocalization where the students read a reminder card with the statement 'I will stop, listen, look, and think before I answer' before each item; Feedback on success and failure; Standard administration.

Subjects Participants included 101 subjects identified as having behavioral and emotional difficulties, with a mean age of 11.2. Of these students, 75 were males and 26 were females. (Children with known neurological dysfunction were not included).

Dependent Variable Two tests were administered: (a) Porteus Maze Test (1942) and (b) Wechsler Intelligence Scale for Children-Revised (WISC-R)

Findings Conditions that provide knowledge of success and those in which payment is given for desired behaviors were found to be powerful motivators for improving the test performance of emotionally disturbed boys and high-impulsive children. Conversely, emotionally disturbed girls and low-impulsive children performed best when given information on the success of their performances.

• Low-impulsive subjects ranked highest on the Feedback and Standard administrative procedures while high-impulsive subjects ranked highest on the Reward for Success and Reinforcing Attention procedures

• For the Full Scale and Performance IQ scores, boys ranked highest on the Reward for Success and the Standard conditions while for the Verbal IQ scores boys ranked highest on the Reward for Success and the Reinforcing Attention conditions
• The Feedback condition ranked highest across IQ scores for girls

Johnson, Bradley-Johnson, McCarthy, & Jamie [51] 1984

Adaptation The test was administered under two conditions (a) Standardized administration, and (b) Standardized administration with token reinforcement.

Subjects *Experiment 1:* Participants included 20 black, elementary-age children who had been classified as educable mentally impaired. The students resided in a rural county and came from families of low socioeconomic status. Mean age was 10-3.

Experiment 2: Participants included 22 black, junior-high age children classified as educable mentally impaired from a mid-eastern city. Ages ranged from 12-7 to 14-11.

Dependent Variable The WISC-R Intelligence Test, except the Mazes subtest, was administered.

Findings *Experiment 1:* The results showed that elementary age, black children evinced WISC-R scores that were an average of nine points higher if they were reinforced with tokens for correct responding rather than receiving only standardized administration procedures. Only the Verbal IQ scores, however, were higher for the experimental group than the control group. No difference was noted for the Performance IQ scores.

Experiment 2: The scores were not significantly affected by token rewards.

Koegel, Koegel, & Smith [53] 1997

Adaptation This experiment employed two different testing conditions (the standardized condition and a motivation/attention condition). In both conditions the examiners verbally encouraged the children and provided verbal and edible rewards contingent upon appropriate test-taking behavior.

Subjects Participants in this study included six pre- and elementary school-aged children, five boys and one girl. All were diagnosed with autism. The children's ages ranged from 3-1 to 9-6.

Dependent Variable Four standardized language and intelligence tests were given: Assessment of Children's Language Comprehension (ACLC; Foster, Giddan, & Stark, 1973) Multiple Components Test Section; ACLC Vocabulary Test Section; Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn & Dunn, 1981); Intelligence Tests.

Findings With only one exception, the test scores for the 44 separate testing sessions were always higher in the motivation/attention condition. The higher test scores under the motivation/attention condition were evident for receptive vocabulary tests, receptive language tests, verbal intelligence tests, and nonverbal intelligence tests. Three children, unable to reach a measurable standard score under the standardized test condition, were sometimes able to score in the normal range when the motivation/attention techniques were implemented.

Saigh & Payne [86] 1979

Adaptation Three reinforcement categories were employed: tokens, verbal praise, and verbal neutral. Subjects in each of these three categories were randomly assigned to two schedule categories, fixed-ratio and continuous reinforcement.

Subjects The study included 120 institutionalized children (equal numbers of males and females) with educable mental retardation. The mean IQ was 65.3, and the mean age was 11.8 years.

Dependent Variable The Arithmetic, Digit Span, Picture Completion, and Block Design subtests of the WISC-R Intelligence Test were administered.

Findings The overall analysis of variance tests for the type of reinforcer groups were significant for the Arithmetic, Digit Span, and Picture Completion subtests. A subsequent Scheffé post-hoc test revealed a significant difference in mean scaled scores for both the verbal praise and token groups relative to the verbal neutral group. There was no main effect for type of reinforcer with the Block Design scores. None of the analyses for a main effect due to level of reinforcement schedule were statistically significant.

Smeets & Striefel [89] 1975

Adaptation Four reinforcement categories were used: end-of-session reinforcement, non-contingent reinforcement, delayed reinforcement, and immediate reinforcement.

Subjects The subjects were 45 deaf and hard-of-hearing children ranging from 11 to 18 years of age. All children had been excluded from regular education programs for the deaf and participated in a special program designed to remedy their academic and behavioral deficits.

Dependent Variable The Raven Progressive Matrices were used for both a pretest and a posttest. Only the posttest involved the four reinforcement categories.

Findings Of the four reinforcement conditions, immediate delivery of checkmarks for correct responses increased test performance the most.

Terrell, Taylor, & Terrell [95] 1978

Adaptation The test was administered under four reinforcement conditions: none, tangible (candy), social, and culturally relevant social.

Subjects The study included 80 second grade students from low socioeconomic backgrounds.

Dependent Variable One measure was used: WISC-R Intelligence Test.

Findings Intelligence was significantly different when students received various reinforcement conditions: Nonreinforcement = Social, Social = Tangible, Tangible > Nonreinforcement, Culturally Sensitive > Nonreinforcement, and Culturally Sensitive > Social

Terrell, Terrell, & Taylor [96] 1980

Adaptation Examinees were randomly assigned to either a Black or White examiner and to one of four reinforcement conditions: none, tangible (candy), social, and culturally relevant/social.

Subjects Participants included 120 Black males aged 9-11 years old.

Dependent Variable The dependent variable was the WISC-R Intelligence Test.

Findings No main effect was found for race of examiner. Significant differences were found for type of reinforcer. An interaction between these variables also was found: Black children performed highest with White examiners using tangible reinforcers or Black examiners using either tangible or culturally relevant reinforcers.

Terrell, Terrell, & Taylor [97] 1981

Adaptation Four reinforcement conditions were implemented: non-reinforcement, candy, social, and culturally relevant/social.

Subjects Participants in this study included 100 black males, aged 9 to 11 years of age. All were enrolled in special education classes under the diagnosis of mild mental retardation.

Dependent Variable The WISC-R Intelligence Test was administered.

Findings Children given tangible or culturally relevant rewards obtained significantly higher scores than did children given either no reinforcement or traditional social reinforcement.

Willis & Shibata [113] 1978

Adaptation The test was administered under three conditions: standard, feedback, and reinforcement (tokens).

Subjects Participants included 30 preschool children (20 boys and 10 girls) ages 3 to 3-6, from lower socio-economic families.

Dependent Variable IQ tests were administered involving the following subtests: Vocabulary, Arithmetic, Picture completion, Geometric design, Information, Similarities, Animal house, Mazes, Block design

Findings Significant differences appeared among the first 5 subtests but no differences were found for them between the two treatment conditions. Seven children in the tangible reinforcement group showed marked changes. Feedback did not result in an increased number of correct responses.

Young, Bradley-Johnson, & Johnson [114] 1982

Adaptation This study utilized three groups: Control group, Delayed reinforcement group, and Immediate reinforcement group.

Subjects Participants included 30 white children with mental retardation, 19 boys and 11 girls. Mean age was 10.5 years. All children came from families of low to middle socioeconomic status.

Dependent Variable Two tests were administered to each child: Slosson Intelligence Test (1963) to assess initial equality between the three groups, WISC-R Intelligence Test.

Findings Both the immediate and delayed reinforcement groups showed significantly better performance than the standardized testing group, although the two reinforcement groups did not differ from one another. Half the children in each of the groups that earned tokens scored above the mentally impaired range while only one child in the standardized testing group scored above the mentally impaired range.

Other: Instruction on Test Taking Strategies

In this area, students are prepared to take the test and given strategies for optimizing their performance.

Background and Foundational Research

As increasingly high stakes are placed on test outcomes, it is very likely that attention turns to the very basic skills needed to simply complete the test as well as providing students with test taking strategies. Such skills include strategies for reading passages, comparing items with each other to ferret out more or less probable answers with or without reference to the passage (referred in the literature as item dependence), eliminating obviously incorrect items, arrangement of the environment, and preparation for the test itself (e.g., sleep, anxiety reduction, etc.). Given the fact that students with disabilities frequently are excluded from statewide testing programs, such differential effects might be expected when providing test taking strategies for students with disabilities. Often, they have had a dearth of experience in formally taking tests. To the degree that such preparation eliminates irrelevant variance from the performance, this focus therefore, may be quite helpful. However, if it limits the generalizations that can be made of performance, it may pose a serious problem.

Analysis of Literature by Subjects and Test

Four studies were found that investigated the effects of test taking and preparation strategies; none very recent. Scruggs, Mastropieri and Tolfa-Veit (1986) reported improved performance on some tests but not others for 4th grade students both with and without learning disabilities when they used a package of test taking strategies with the Stanford Achievement Test. For Rogers and Bateson (1991), the focus was on test

wiseness of 12th grade students taking high school exit examinations in several different subject areas. Their conclusion was that partial knowledge improved performance and students who were test wise attained higher scores. None of their students was identified as having a disability. Rather than passively analyze how 'wise' students are in taking tests, Rozinski and Bassett (1992) actually coached college students on how to eliminate options within multiple choice items on the Scholastic Aptitude Test. They were successful on many items, particularly analogy items. Finally, McAuliffe (1993) attempted to heighten "deeper understanding and attention" in eighth grade students who were at risk students of failing reading and writing tasks. They reported better performance on later samples compared to earlier ones with classroom tasks. However, they also documented the loss of such strategy use with a state administered test, where students primarily focused on "getting the right answer." Finally, Whinnery and Fuchs (1993) reported positive effects from teaching students how to take tests by reviewing correct items, answering problems according to perceived difficulty, and using a specific goal strategy.

Analysis of Research Quality and Summary

This research area is in need of better research designs, with random assignment of students to treatments and clearer definitions of treatments that isolate specific skills. Most of this research has included intact groups, used quasi-experimental designs, and implemented treatment packages. The research also has not focused on students with and without disabilities to determine differential effects.

Annotated References of Investigations on Other:

Instruction on Test Taking Strategies

McAuliffe [63] 1993

Adaptation Instructional planning was analyzed to increase student interest and background knowledge, interpersonal negotiations were conducted to focus encourage students to construct more elaborated meaning, reading and discussion was used to promote reflection, prediction and inferences, sharing was used to expose students to each others' comprehension process, high interest themes were developed to engage students, and finally, practicing for the test was used to accentuate differences in classroom and test contexts.

Subjects An eighth grade class participated in this study. A total of 15 students were judged to be at-risk readers.

Dependent Variable Early samples of reading and writing are compared with later samples. A state-mandated reading test also was analyzed.

Findings Differences are highlighted between the more "authentic" assessment of the classroom activities and the surface activities of the test administration context: Students were actively involved in more "authentic" literary processes as they negotiated meaning in the supportive instructional context; Practicing for the Illinois Reading Assessment appeared to move students away from the empowered stances they developed during instruction; During assessment practice students did not seem to be trying to figure out the text but seemed to be trying to choose the "right" answer.

Rogers & Bateson [84] 1991

Adaptation A post-hoc analysis of test wiseness was conducted, coding strategies into the following groups: (a) Eliminate options known to be incorrect, (b) Choose neither or one of two options if one being correct implies the other is incorrect, (c) Choose either none or one option if being correct implies other is correct, (d) Select option consistent with stem, and (e) Use relevant content information in other items

Subjects The participants were 954 twelfth grade students who wrote provincial examinations for English 12 in order to graduate. The students were from 10 public schools in British Columbia.

Dependent Variable Provincial examinations: English, Algebra, Geography, History, Biology, Chemistry

Findings Upwards of 43% to 80% of test items can be answered by test-wise students. Test wiseness serves as an enhancer of performance when students have partial knowledge. Students with partial knowledge and test wiseness will perform better than students with only one of these attributes.

Roznowski & Bassett [85] 1992

Adaptation The focus was on 'coachability' of items to help examinees eliminate options. Three conditions were studied with review sheets provided in the last two conditions: (a) Control condition with no coaching, (b) 1 hour of encouragement with students given general motivation and coaching, and (c) 1 hour of training on taking analogy tests like looking for particular kinds of relationships among word pairs with no or partial knowledge of stem words.

Subjects Participants included 100 undergraduate psychology students randomly assigned to one of the three conditions.

Dependent Variable The following Scholastic Aptitude Test analogy subtests were administered: Analogy (10 items), Antonyms, Sentence completion, Reading comprehension

Findings Results indicated:

- Five of 10 analogy items were made significantly easier with coaching.
- Proportion correct indices were higher with students who were coached.

Scruggs, Mastropieri, & Tolfa-Veit [88] 1986

Adaptation Students were trained in test taking behaviors in reading and math in five sessions: (a) Use of thinking strategies in reading sub-tests, (b) Attending to appropriate cues, (c) Use of test materials, (d) Location and review of information, (e) Double checking answers, (f) Self-monitoring strategies, and (g) Work completion strategies.

Subjects Participants included 85 students in grade 4 with 44 students with learning disabilities. There were 63 boys and 22 girls.

Dependent Variable Several areas of the Stanford Achievement Test were administered: Reading Comprehension, Word study, Math concepts, Math computation, Math applications

Findings Three findings were reported:

- Students with behavior disorders performed as well as students with learning disabilities.
- On word study skills and math concepts, training improved performance.
- On reading comprehension and math applications, training did not improve performance.

Whinnery & Fuchs [112] 1993

Adaptation Students were taught to use a test taking strategy that involved: (a) Reviewing correct items on their most recently completed test, (b) Answering problems on the current test based on level of perceived difficulty, (c) A goal strategy was also used.

Subjects The participants were 40 students with mild learning disabilities, grades 2 –8.

Dependent Variable Curriculum-based measures (CBM) were used to track student progress toward arithmetic computation goals.

Findings Students who received the CBM test-taking strategy training scored higher on a posttreatment computation test than students with no CBM test-taking strategy training.

Other: Instructional Level Testing

Many tests establish basal and ceiling levels so that most items are within a range of difficulty that is appropriate for the student. When such levels are not established, the standard error of measurement is much greater, especially at performance levels that are further from the mean. This accommodation, also referred to as out-of-level testing, involves the student being tested at an instructional rather than grade level and often results in items from an earlier grade level form being administered.

Analysis of Literature by Subjects and Test

When taking an instructional level test versus a grade level test, Long, Schaffran, and Kellogg (1977) reported improved performance for 2nd and 3rd graders taking the Gates-McGinitie (reading) test but decreased performance for 4th graders. At all grade levels, different decisions on Title I eligibility and program successes were affected by the level of testing.

Analysis of Research Quality and Summary

Obviously, with only one study conducted, definitive conclusions are hardly appropriate. As computer assisted testing becomes increasingly used, more research on instructional level is likely to be forthcoming, given that this type of testing focuses on item presentation at the student's instructional level. Additionally, significant concerns exist about the technical adequacy of the data generated from out-of-level testing.

Annotated References of Investigations on Other:

Instructional Level Testing

Long, Schaffran, & Kellogg [58] 1977

Adaptation. The students were tested at two levels, their grade-level and their instructional level. Half of the students received the grade level form first and half received the instructional level form first.

Subjects A total of 482 students were selected in grades two, three, and four who were determined to have an instructional level at least one grade level below their actual grade level as determined by the Botel Word Opposites Test.

Dependent Variable The Gates-McGinitie series administered in this study were: Primary A, grade1; Primary B, grade2; Primary C, grade 3; Survey D, grades 4-6. Instructional level testing provided higher grade equivalent scores for grades two and three, and lower grade equivalent scores for grade four, than did grade level testing.

Findings It was found that grade level testing identified larger numbers of students eligible for Title I program assistance at grades two and three. Instructional level testing, however identified larger numbers of students at grade four.

- At every grade, and on both subtests, instructional level testing resulted in more students reaching the criterion.
- Grade equivalent score means were found to differ, numbers of students eligible for reading programs differed, and measures of student and program success varied, depending upon whether grade-level or instructional level tests were administered.

CRITICAL QUESTIONS TO ADDRESS IN TEST CHANGE RESEARCH

A series of questions are posed to help structure the review on test changes and help establish an empirical basis for determining whether those test changes are accommodations or modifications. Rather than simply list findings from the past two decades of research, these questions help structure the inquiry process and move findings toward conclusions. The first three questions focus broadly on the research from a documentary perspective, taking an inventory of the investigations and reflecting on the application of the findings to practice

based on the work. After considering the applicability, the next three questions address the quality of the research both experimentally and theoretically. Finally, attention is given to the consequences of changing testing practices in our nation's schools and the need to ensure that teachers have an adequate knowledge base to make decisions about what types of changes to make and for whom.

All of these issues are related: in the generalizations of the results, in the assurance that those results are valid, and in the Individualized Education Program (IEP) teams' decision-making system likely to result. If students with disabilities have difficulty accessing the test because of the manner in which the test is administered or taken, problems arise in the validity of inferences that can be made. While many accommodations are being implemented in the absence of empirical data (Anderson, Jenkins, & Miller, 1995; Chin-Chance, Gronna, & Jenkins, 1996; Shriner, Gilman, Thurlow, & Ysseldyke, 1995), this review hopefully provides enough findings to determine which changes in testing are supported.

Because data obtained from large-scale assessments are used to make many different kinds of decisions (Ysseldyke, Thurlow, McGrew, & Vanderwood, 1994) and given the high-stakes context of most statewide assessment systems, the general call from researchers and reformers alike is to provide an evaluation of the "qualities of validity, reliability, and fairness" (NCEST, 1992, p. 27). As Linn (1993) noted "If the NCEST provision for obtaining supporting validity evidence is taken seriously, then validators will have a full agenda" (p. 6).

Addressing validity as a quantifiable concept has become more complex over the last two decades. No longer is the psychometric view of validity dominant, packaged neatly as content, criterion-concurrent, criterion-predictive, and construct. Messick (1989a,

1989b, 1995a, 1995b) has proposed a more unified version in which these and other types of evidence and consequences are considered. Furthermore, validity itself has evolved from a purely measurement issue to a political and social issue.

> Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions but also of the persons responding as well as the context of the assessment...These issues are critical for performance assessment because validity, reliability, comparability, and fairness are not just measurement principles; they are social values that have meaning and force whenever evaluative judgments and decisions are made. As a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation between test scores and purported criterion (i.e., classical criterion validity) or by expert judgments that test content is relevant to the proposed test use (i.e., traditional content validity) (Messick, 1995a, p.5).

Are the Findings Relevant for Classroom Practice and Instructional Focus?

This question considers the degree to which the outcomes from research on test changes are anchored in the classroom and useable by teachers. This question speaks to the relevance of findings to state guidelines for test changes. What populations are being studied in our test change research, and what types of tests are being changed?

Whom Have We Studied?

To answer this question, we need to determine if the subjects in the research are comparable to the students in our schools. If we are to use the results from this research, we need to be certain that the studies have external validity. Have we sampled students appropriately so that the findings can be generalized to others? What age groups and disability areas have we included in the research?

In general, a range of students from diverse backgrounds and of varying ages has participated in our research. It is likely that the research based on the broadest (most diverse) group is most generalizable and that if consistent target groups have been identified, the findings need to be limited to comparable students.

For example, the research on extended time has covered a full range of ages, from elementary to college age. In contrast, the research on examiner familiarity has been confined to young children, often of preschool ages. Can we extrapolate the findings

from the extended time or examiner familiarity to all students? In this instance, the findings of extended time but not examiner familiarity would apply more broadly to most large-scale testing situations that use high power tests. Nevertheless, examiner familiarity may be an important issue for young students with little experience in taking tests.

Another example would be in the use of computers in the writing process. Generally, participating students have been at least middle school age and older and often in college. From a common sense point of view, students need to have keyboarding skills to make the efficient use of computers a reasonable test change. And it is not until grades 6 and later that such training is formally built into the course sequences of most schools. And yet, by grade 12, most writing classes tend to use computers and word processors within the teaching-learning cycle. Therefore, the findings from this research may need to be limited to this older age group and in reference to the practices in appropriate settings.

What Tests Have Been Used to Study Changes and For Which Decisions?

Not only have students been from very diverse backgrounds and with varying ages and disabilities, but the kinds of tests being changed have been remarkably different from each other, covering intelligence tests, broad achievement tests, and specific skill tests. No pattern exists, however, in the findings, this reflects the fact that some tests as more or less robust in their capacity to be "validly" changed. For example, intelligence has generally been assumed to be a relatively stable "trait" and one would assume that, although it should be given in a standardized fashion, some (superficial) changes really should not have a major impact. Yet, the research on reinforcement indicates that students' performance is substantially affected with different types and schedules. In general, few published norm-referenced tests have been used while many admissions tests have been used. Other measures have typically been researcher-defined using established tests in experimental contexts of a study. Increasingly, state tests themselves are being evaluated as in the case of the research on Kentucky's test.

Tests have specific task demands and therefore delimit the comparability-generalizability of the findings (Linn, 1993; Miller, 1996; Ruiz-Primo, Baxter, & Shavelson 1993). Likewise, the technical adequacy, reliability, and certification of tests needs to be considered (LeMahieu, Gitomer, & Eresh, 1995; Linn, 1993; Linn, Baker,

Dunbar, 1991; & Reckase, 1995), particularly as they pertain to special education populations (Poteet, Choate, & Stewart, 1993; Tindal, 1997). In addition, any examination of assessment systems also requires their validation in a decision-making framework (Messick, 1995; Winter, 1996).

How Well Designed is the Research on Test Changes and Can the Results be Trusted?

This question is essentially methodological and focuses on the way in which we conduct research. In many instances, the research is based on using intact groups with post-hoc evaluations of outcomes. Occasionally, actual experimental designs are used with students randomly assigned to treatments in a crossed or nested manner. What is the difference in the outcomes generated by these designs and how accurate are conclusions in identifying cause-effect relationships?

Have We Done our Research Correctly (with Reliability and Validity)?

This question addresses the internal validity of the research that has been conducted, with threats to validity controlled or eliminated. Such control allows stronger interpretations to be made about cause and effect relationships. Following is an illustrative list of threats that have been reported in the professional literature on research design.

1. Have the studies been conducted so historical events have not occurred while the data were being collected?
2. Has the study been done in a timely fashion so students have not grown older during the data collection process with developmental changes influencing performance?
3. Has the testing process itself been implemented in a manner that precludes students from reacting and influencing their performance?
4. How have students been selected for the study and assigned to treatments?
5. Have students been assigned to treatment groups so that they do not know that they are in a "control" group or do not compete with each other or across conditions?
6. Have the testing instruments been consistent in the manner in which they have been calibrated?

7. Is the test change adequately described in the study and been confined to one element, avoiding multiple changes from being implemented concurrently.
8. Have judges been unbiased in their evaluations so that they use only the target criteria and not other criteria in scoring or judging performance?
9. Have all of these issues been avoided singly or in any combination during the time that test changes have been monitored?

All of these issues affect whether we can attribute a cause-effect relationship between independent and dependent variables. Also, included in the threats to validity is statistical conclusion validity that addresses the measurement and analysis of outcomes. Have the data been analyzed appropriately and with the correct statistical test?

In general, the research on test changes is perched in a fragile position between program evaluation and quasi-experimental research. Although data from program evaluation may be quite accurate and correct in depicting outcomes, the findings often are of limited utility in explaining those outcomes. In contrast, while quasi-experiments help establish cause-effect explanations, they come with enough threats to validity controlled to warrant systematic replications. Probably the most significant problem with the research on test change is the lack of clear identifying independent variables that isolate a specific change. Rather, most of the literature includes many changes in testing that occur concurrently. And while many changes have been validated, this research primarily is based on group designs, making it impossible to predict the effect of any specific test change for an individual student.

Does the Research on Test Changes Help Establish Construct Validity?

It is critical that research helps educators understand what is meant when changes in testing are implemented. Why are changes in testing needed, and what is their effect on interpretations of performance? Do test changes eliminate irrelevant access skills, level the playing field, or avoid unfair advantage and, if so, how? What do these phrases mean? Have the changes in testing been studied in the context of the teaching-learning cycle, reflecting systemic validity as Fredericksen and Collins (1989) described the interrelationship between measures of achievement and attempts to influence it.

The validity of test changes (whether to support them as an accommodation or as a modification that results in something else being measured) rests on three assumptions:

1. The changes that do not alter the construct of the measure;
2. The changes are based on individual need; and
3. The outcomes produce differential effects (i.e., work with those who need it and not for those who do not need it).

To the degree that these three attributes are not all in concert, the change becomes more than an accommodation it becomes a modification which affects further decision making.

Construct of the measure. All tests and measures need to be defined in terms of purpose, objectives, and domain sampling. Salvia and Ysseldyke (1998) define 13 purposes for testing, some of which address entitlement decisions and others instructional evaluation, both formative and evaluative, as well as accountability decisions. For state education agencies, tests generally are given to provide an evaluation of schools and schooling. In some states, this accountability is at the student level, with students receiving individual certificates, diplomas, or awards as a function of performance on a test. In other states, the individual being judged is the teacher with aggregate levels of performance from their classroom used to make judgment and the focus is instructional. Finally, school systems at the building or district level are being evaluated as part of a larger accountability system. At all levels of evidence and decision making, the construct being measured must be defined in terms of content and sampling plans. Content is typically defined by reference to objectives listed in curriculum frameworks or learning outcomes while sampling plans provide rules describing who gets tested with which forms and items.

Individual need. Large-scale testing programs usually are designed to provide comparable opportunity for all students to perform. The focus is on comparability and all students. To achieve this outcome, typically, decisions are made about the range of changes that are possible and do not appear to threaten the construct of the measure; these considerations often are balanced by individual need. This criterion, however, is not to be confused with optimal performance. In many cases, changes in testing may in fact improve performance but the goal of most large-scale testing programs is to describe not *best* but *typical* performance.

Therefore, opportunity-to-learn has become an important issue. Measurement of content and performance standards must be accompanied with assessment of opportunity-

to-learn and equity standards. Both of these dimensions need to be empirically evaluated through the use of educational indicators that reflect specific opportunities (Oden, 1990). This linkage may eventually need to entail analyzing instructional coverage (Linn, 1983; Miller & Linn, 1988; Tindal & Nolet, 1996) and documenting content collaboration between general and special education (Nolet & Tindal, 1996; Tindal & Nolet, 1996). Finally, school delivery standards (Porter, 1993, 1995) may need to include macro variables like adequacy of resources, schedules, staff, and professional development activities.

Differential outcomes. If a change in testing works for all or fails to work for all, then this change is group defined, not individually defined. Essentially, this attribute focuses on an interaction between the type of student and the type of change. While students can be classified according to disability, it is probably better to think of grouping them according to level or type or assistance using the IEP to document this decision.

It is unquestionably easier to demonstrate a test change is an accommodation for a person with a physical disability and does not alter the construct being tested than to prove the lack of nexus between an accommodation for a mental disability and the construct being tested (Phillips, 1994). Hartman and Redden (1985) concur by asserting that the purpose of an accommodation is not to "give disabled students a competitive edge, but rather eliminate competitive disadvantage" (p. 2). Knowledge of appropriate test changes for students with disabilities, however, originates from "measurement competent educators" (O'Sullivan & Chalnick, 1991, p. 17). Therefore, the change in tests must be validated with findings of an interaction between students with and without disabilities as they perform with and without the change: Students with disabilities would perform higher with the accommodation while no such changes would be found for students without disabilities.

When Research is Put into Practice, What are the Consequences at a System Level?

How do we systemically establish policy and employ practice on the basis of the research findings? What are the intended and unintended outcomes and social consequences from the decisions to change tests in an effort to accommodate students with disabilities?

How is the entire educational system affected and for which subgroups of students are the effects pronounced and helpful or harmful? In the end, we believe the results of this research need to be embedded in practice consistently by teachers.

State Practices and Teacher Knowledge: What Next?

Because appropriate decisions about high stakes tests must be made for students with disabilities, teachers are expected to function as "measurement competent educators" who can "evaluate student performance in a fair and meaningful way" (Siskind, 1993a, p. 233)? All teachers, therefore, must be knowledgeable about assessment and assessment-related concepts. However, important deficits in teachers' knowledge concerning high-stakes testing often are evident. Most of teachers' knowledge about testing and measurement comes from "trial-and-error learning in the classroom" (Wise, Lukin, & Roos, 1991, p. 39). Furthermore, knowledge deficits are not specific to general education teachers. Shepard's (1983) research documents that even school psychologists were likely to lack competence in their knowledge and application of assessment. Corroborating Shepard's conclusions, Siskind (1993b) and Hollenbeck, Tindal & Almond (1998) reported that special educators are not well informed about assessment and assessment procedures either.

Wise et al. (1991) attributed this lack of assessment knowledge to the fact that teacher certification agencies at the state-level that do not require assessment or measurement courses for initial teacher certification. The findings of Wise et al. (1991) are also supported by Schafer's (1991) review of research. "Only about half of the teacher education programs in the nation require a course in measurement for initial certification" (p. 3). This research is further confirmed by Stiggins (1991), when he reports that "less than half [of the major teacher training institutions in the six western states] offered any assessment training at all" (p. 7).

This current lack of assessment knowledge at the teacher level manifests itself as "test score pollution" or variations in test administration that increase the error component of the outcomes. As Nolen et al. (1992) conclude, "administration practices ranged from the innocuous to the clearly unethical" (p. 13). This variation in administration on a state test was further documented by Hollenbeck et al.'s (in 1998) research on teachers' knowledge of appropriate accommodations. Likewise, the NCEO (1996) has reported that a

comprehensive set of empirical research on testing accommodations simply does not exist. Therefore, teacher knowledge, perceptions, and acceptability of test changes (Gajria, Salend, & Hemrick, 1994; Jayanthi, Epstein, Polloway, & Bursuck, 1996; Miller, 1996; Siskind, 1993) must be improved. Otherwise, the field cannot understand the issues in participation rates, evaluate state and district policies for implementation of accommodations, implement test changes consistently across schools and districts, or determine statistical comparability of various test changes.

This review of test changes represents the first step in improving the knowledge base. As state directors of testing make decisions about which changes are allowable and which are not allowable, the studies in this document should be referenced both as a primary reference and a synthesized review. Decisions could then be made in a timely and accurate manner. Ideally, over time, the field should become more sophisticated about this decision-making process.

With a more knowledgeable educational workforce, it might be possible to require not only that IEPs become more functional and appropriate but also that the decision-making system for recommending test changes become more systematic. Rather than simply requiring teams to make the best decision they can, state directors of testing and special education directors could provide a more prescriptive and empirical approach that is based on what we know so far. A more clear distinction would exist then between changes that are appropriate accommodations versus those that reflect substantial changes and therefore are modifications.

The final recommendation from this report is that the research process for creating policy becomes more anchored to an experimental rather than descriptive or comparative approach. In setting the research agenda, however, it is clear that many more changes are being recommended in practice than we have data to support. Somehow, researchers and practitioners need to collaborate more effectively and conduct broader research on various test changes. This research however needs to be framed appropriately and executed carefully, on more diverse student populations, different tests, and with different decisions.

References from Research on Test Accommodations

1. Abikoff, H., Courtney, M. E., Szeibel, P. J., & Koplewicz, H. S. (1996). The effects of auditory stimulation on the arithmetic performance of children with ADHD and non-disabled children. *Journal of Learning Disabilities, 29*, 238-246.
2. Alster, E. H. (1997). The effects of extended time on algebra test scores for college students with and without learning disabilities. *Journal of Learning Disabilities, 30,* 222-227.
3. Ansley, T. N., & Forsyth, R. A. (1990). An investigation of the nature of the interaction of reading and computational abilities in solving mathematical word problems. *Applied Measurement, 3*, 319-329.
4. Arick, J. & Nave, G. (1997). *A Full Evaluation Study of the Oregon Supported Education Plan and Its Impact Upon Student Outcomes. Final Report* (Program Description). Portland State University, Oregon School of Education.
5. Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written vs. word-processed papers* (Report No. 143). Whittier, CA: Rio Hondo College. (ERIC Document Reproduction Service No. ED 345 818)
6. Baxter, B. (1931). An experimental analysis of the contributions of speed and level in an intelligence test. *The Journal of Educational Psychology, 22*, 285-296.
7. Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modifications on the minimum competency performance of learning disabled students. *Learning Disability Quarterly, 6*, 75-77.
8. Bennett, R. E., Rock, D. A., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *The Journal of Special Education, 21* (3), 9-21.
9. Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24* (1), 44-55.
10. Braden, J. P., Elliott, S. N., & Kratochwill, T. R. (1997). *The performance of students with and without exceptional educational needs on performance assessment and multiple choice achievement measures.* Paper presented at the

Council of Chief State School Officers National Conference on Large Scale Assessment, Colorado Springs, CO.

11. Bradley-Johnson, S., Graham, D. P., & Johnson, C. M. (1986). Token reinforcement on WISC-R performance for white, low-socioeconomic, upper and lower elementary-school-age students. *Journal of School Psychology, 24*, 73-79.
12. Bradley-Johnson, S., Johnson, C. M., Shanahan, R. H., Rickert, V. I., & Tardona, D. R. (1984). Effects of token reinforcement on WISC-R performance of black and white, low socioeconomic second graders. *Behavioral Assessment, 6*, 365-373.
13. Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement, 32*, 323-340.
14. Burk, M. (1998, October). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities.* Paper presented at Office of Special Education Program Cross Project Meeting "Technology and the Education of Children with Disabilities: Steppingstones to the 21st Century.
15. Centra, J. A., (1986). Handicapped student performance on the Scholastic Aptitude Test. *Journal of Learning Disabilities, 19*, 324-327.
16. Cohen, A. S., & Kim, S. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education, 5*, 303-320.
17. Coleman, P. J. (1990). Exploring visually handicapped children's understanding of length (math concepts). (Doctoral dissertation, The Florida State University, 1990). *Dissertation Abstracts International, 51*, 0071.
18. Curtis, H. A., & Kropp, R. P. (1961). A comparison of scores obtained by administering a test normally and visually. *Journal of Experimental Education, 29*, 249-260.
19. Dalton, B., Morocco, C. C., Tivnan, T., & Rawson, P. (1994). Effect of format on learning disabled and non-learning disabled students' performance on a hands-on science assessment. *International Journal of Educational Research, 21*, 299-316.
20. Dalton, B., Tivnan, T., Riley, M. K., Rawson, P., & Dias, D. (1995). Revealing competence: Fourth-grade students with and without learning disabilities show what

they know on paper-and-pencil and hands-on performance assessments. *Learning Disabilities Research & Practice, 10*, 198-214.

21. Derr-Minneci, T. F. (1990). A behavioral evaluation of curriculum-based assessment for reading: Tester, setting, and task demand effects on high- vs. average- vs. low-level readers (high-level readers, average-level readers) (Doctoral dissertation, Lehigh University, 1990). *Dissertation Abstracts International, 51*, 0105.
22. Elliott, S. N., & Kratochwill, T. R. (1998a). *Experimental analysis of the effects of testing accommodations on the scores of students with disabilities*. Unpublished manuscript, University of Wisconsin-Madison.
23. Elliott, S. N., & Kratochwill, T. R. (1998b). *Performance assessment and standardized testing for students with disabilities: Psychometric issues, accommodation procedures, and outcome analysis.* Unpublished manuscript, University of Wisconsin-Madison.
24. Espin, C. A., & Sindelar, P. T. (1988). Auditory feedback and writing: Learning disabled and nondisabled students. *Exceptional Children, 55*, 45-51.
25. Fuchs, D., Fuchs, L.S., Garwick, E.R., & Featherstone, N. (1983). Test performance of language-handicapped children with familiar and unfamiliar examiners. *The Journal of Psychology, 114*, 37-46.
26. Fuchs, D., Featherstone, N. L., Garwick, D. R., & Fuchs, L. S. (1981). *The importance of situational factors and task demands to handicapped children's test performance* (Research Report No. 54). Institute for Research on Learning Disabilities: University of Minnesota
27. Fuchs, D., Featherstone, N. L., Garwick, D. R., & Fuchs, L. S. (1984). Effects of examiner familiarity and task characteristics on speech-and-language-impaired children's test performance. *Measurement and Evaluation in Guidance, 16* (4), 198-204.
28. Fuchs, D., & Fuchs L. S. (1989). Effects of examiner familiarity on Black, Caucasian, and Hispanic children: A meta-analysis. *Exceptional Children, 55*, 303-308.

29. Fuchs, D., Fuchs, L. S., Dailey, A. M., & Power, M. H. (1985). The effect of examiners' personal familiarity and professional experience on handicapped children's test performance. *Journal of Educational Research, 78* (3), 141-146.

30. Fuchs, D., Fuchs, L. S., Garwick, D. R., & Featherstone, N. (1983). Test performance of language-handicapped children with familiar and unfamiliar examiners. *The Journal of Psychology, 114*, 37-46.

31. Fuchs, L.S., Fuchs, D., Karns, K., Hamlett, C., Katzaroff, M., & Dutka, S. (1998). Comparisons among individual and cooperative performance assessments and other measures of mathematics competence. *Elementary School Journal, 98*, 23-30.

32. Fuchs, D., Fuchs, L. S., & Power, M. H. (1987). Effects of examiner familiarity on LD and MR students' language performance. *Remedial and Special Education, 8* (4), 47-52.

33. Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (in press). *Supplementing teacher judgments about test accommodations with objective data sources.* School of Psychology Review.

34. Gallina, N. B. (1989). Tourette's syndrome children: Significant achievement and social behavior variables (Tourette's syndrome, attention deficit hyperactivity disorder) (Doctoral dissertation, City University of New York, 1989). *Dissertation Abstracts International, 50*, 0046.

35. Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76* (1), 35-40.

36. Halla, J. W. (1988). A psychological study of psychometric differences in Graduate Record Examinations General Test scores between learning disabled and non-learning disabled adults (Doctroal dissertation, Texas Tech University, 1988). *Dissertation Abstracts International, 49*, 0230.

37. Harker, J. K., & Feldt, L. S. (1993). A comparison of achievement test performance of nondisabled students under silent reading and reading plus listening modes of administration. *Applied Measurement, 6*, 307-320.

38. Harris, G. S. (1992). *Assessing problem-solving skills on selected questions from the Scholastic Aptitude Test.* Unpublished doctoral dissertation, Rutgers the State University of New Jersey, New Brunswick.

39. Hasselbring, T. S., & Crossland, C. L. (1982). Application of microcomputer technology to spelling assessment of learning disabled students. *Learning Disability Quarterly, 5*, 80-82.

40. Helwig, R., Stieber, S., Tindal, G., Hollenbeck, K., Heath, B., & Almond, P. (1998). *A comparison of factor analyses of handwritten and word-processed of middle school students.* Manuscript submitted for publication, University of Oregon.

41. Helwig, R., Tedesco, M., Heath, B., Tindal, G., & Almond, P. (in press). *The relationship between reading ability and performance on a video accommodated math problem-solving task. Journal of Educational Research.*

42. Hidi, S. E., & Hildyard, A. (1983). The comparison of oral and written productions in two discourse types. *Discourse Processes, 6*, 91-105.

43. Higgins, E. L., & Raskind, M. H. (1995). Compensatory effectiveness of speech recognition on the written composition performance of postsecondary students with learning disabilities. *Learning Disability Quarterly, 18*, 407-418.

44. Hill, G. A. (1984). Learning disabled college students: The assessment of academic aptitude (Doctoral dissertation, Texas Tech University, 1984). *Dissertation Abstracts International, 46*, 0230.

45. Hoffman, D. I., & Lundberg, G. D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational and Psychological Measurement, 36*, 791-809.

46. Hollenbeck, K., Tindal, G., Harniss, M., & Almond, P. (1998). *The influence of computer screen size on statewide writing test scores: answers to an accommodation issue.* Manuscript submitted for publication, University of Oregon.

47. Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1998). *Handwritten versus word processed statewide compositions: Do judges rate them differently?* Manuscript submitted for publication, University of Oregon.

48. Horton, S. V., & Lovitt, T. C. (1994). A comparison of two methods of administering group reading inventories to diverse learners. *Remedial and Special Education, 15*, 378-390.

49. Jackson, A. M., Farley, G. K., Zimet, S. G., & Gottman, J. M. (1979). Optimizing the WISC-R performance of low- and high-impulsive emotionally disturbed children. *Journal of Learning Disabilities, 12* (9), 56-59.

50. Jensen, H. K. (1997). *Differences in reading comprehension between college students with learning disabilities and college students without learning disabilities on the Nelson-Denny Reading Test as related to question type and length of test.* Unpublished doctoral dissertation, The University of North Dakota, Grand Forks.

51. Johnson, C. M., Bradley-Johnson, S., McCarthy, R., & Jamie, M. (1984). Token reinforcement during WISC-R administration II. Effects on mildly retarded black students. *Applied Research in Mental Retardation, 5*, 43-53.

52. Keene, S., & Davey, B. (1987). Effects of computer-presented text on LD adolescents' reading behaviors. *Learning Disability Quarterly, 10*, 283-290.

53. Koegel, L. K., Koegel, R. L., & Smith, A. (1997). Variables related to differences in standardized test outcomes for children with autism. *Journal of Autism and Developmental Disorders, 27*, 233-243.

54. Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: Center for Research on Standards and Student Testing.

55. Lee, J. A., Moreno, K.E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46*, 467-474.

56. Legg, S. M., & Buhr, DC (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11*, 23-27.

57. Linder, E. A. (1989). *Learning disabled college students: A psychological assessment of scholastic aptitude.* Unpublished doctoral dissertation, Texas Tech University, Lubbock.

58. Long, J. V., Schaffran, J. A., & Kellogg, T. M. (1977). Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. *Journal of Educational Measurement, 14*, 203-213.

59. Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*, 31-50.

60. Loyd, B. H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education, 4*, 11-22.

61. Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement, 31*, 251-263.

62. MacArthur, C. A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of Special Education, 21* (3), 22-42.

63. McAuliffe, S. (1993). A study of the differences between instructional practice and test preparation. *Journal of Reading, 36*, 524-530.

64. Mick, L. B. (1989). Measurement effects of modifications in minimum competency test formats for exceptional students. *Measurement and Evaluation in Counseling and Development, 22*, 31-36.

65. Miller, P. (1990). *Use of the Peabody Picture Vocabulary Test-Revised (PPVT- R) with individuals with severe speech and motor impairment: Effect of response mode on test results (speech impairment).* Unpublished doctoral dissertation, University of Kansas, Kansas City.

66. Miller, S. (1998). *The relationship between language simplification of math word problems and performance for students with disabilities.* Unpublished master's project, University of Oregon, Eugene, OR.

67. Mollenkopf, W. G. (1960). Time limits and the behavior of test takers. *Educational and Psychological Measurement, 20*, 223-230.

68. Montani, T. O. (1995). Calculation skills of third-grade children with mathematics and reading difficulties (learning disabilities) (Doctoral dissertation, Rutgers the State University of New Jersey, 1995). *Dissertation Abstracts International, 56*, 0910.

69. Munger, G. F., & Lloyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research, 85* (1), 53-57.

70. Murray, E. A. (1987). The Relationship Between Spatial Abilities and Mathematics Achievement in Normal and Learning-Disabled Boys (Doctoral dissertation, Boston University, 1987). *Dissertation Abstracts International, 49*, 0017.

71. Myers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17*, 347-352.

72. Ofiesh, N. S. (1997). Using processing speed tests to predict the benefit of extended test time for university students with learning disabilities (Doctoral dissertation, The Pennsylvania State University, 1997). *Dissertation Abstracts International, 58*, 0176.

73. Olson, J., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress.* Washington, D.C.: U. S. Department of Education, National Center for Education Statistics.

74. Olswang, L. B., & Carpenter, R. L. (1978). Elicitor effects on the language obtained from young language-impaired children. *Journal of Speech and Hearing Disorders, 42*, 76-88.

75. Perez, J. V. (1980). Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation (Doctoral dissertation, University of South Florida, 1980). *Dissertation Abstracts International, 41*, 0206.

76. Perlman, C. L., Borger, J., Collins, C. B., Elenbogen, J. C., & Wood, J. (1996). *The effect of extended time limits on learning disabled students' scores on standardized reading tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

77. Peterson, R. S. (1998). *Question redistribution as a reading accommodation in statewide assessments.* Unpublished master's project, University of Oregon, Eugene, OR.

78. Pomplun, M. (1996). Cooperative groups: Alternative assessment for students with disabilities. *The Journal of Special Education, 30* (1), 1-17.

79. Powers, D. E., & Fowles, M. E. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement, 33*, 433-452.

80. Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220-233.

81. Raskind, M. H., & Higgins, E. (1995). Effects of speech synthesis on the proofreading efficiency of postsecondary students with learning disabilities. *Learning Disability Quarterly, 18*, 141-158.

82. Rock, D. A., Bennett, R. E. & Jirele, T. (1988). Factor structure of the Graduate Record Examinations General Test in handicapped and nonhandicapped groups. *Journal of Applied Psychology, 73* (3), 383-392.

83. Rogers, W. T. (1983). Use of separate answer sheets with hearing impaired and deaf school age students. *B.C. Journal of Special Education, 7* (1), 63-72.

84. Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*, 159-183.

85. Roznowski, M., & Bassett, J. (1992). Training test-wiseness and flawed item types. *Applied Measurement in Education, 5*, 35-48.

86. Saigh, P. A., & Payne, D. A. (1979). The effects of type of reinforcer and reinforcement schedule on performances of EMR students on four selected subtests of the WISC-R. *Psychology in the Schools, 16*, 106-110.

87. Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. *Educational Assessment, 2*, 325-338.

88. Scruggs, T. E., Mastropieri, M. A., & Tolfa-Veit, D. (1986). The effects of coaching on the standardized test performance of learning disabled and behaviorally disordered students. *RASE, 7*, 37-41.

89. Smeets, P. M. & Striefel, S. (1975). The effects of different reinforcement conditions on the test performance of multihandicapped deaf children. *Journal of Applied Behavior Analysis, 8*, 83-89.

90. Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7*, 211-222.

91. Stoneman, Z., & Gibson, S. (1978). Situational influences on assessment performance. *Exceptional Children, 44*, 166-169.

92. Supovitz, J. A., & Brennan, R. T. (1997). Mirror, mirror on the wall, which is the fairest test of all? An examination of the equability of portfolio assessment relative to standardized tests. *Harvard Educational Review, 67*, 472-506.

93. Swain, C. R. (1997). A comparison of a computer-administered test and a paper and pencil test using normally achieving and mathematically disabled young children (Doctoral dissertation, University of North Texas, 1997). *Dissertation Abstracts International, 58*, 0158.

94. Tachibana, K. K. (1986). Standardized testing modifications for learning disabled college students in Florida (modality) (Doctoral dissertation, University of Miami, 1986). *Dissertation Abstracts International, 47*, 0125.

95. Terrell, F., Taylor, J., & Terrell, S. L. (1978). Effects of type of social reinforcement on the intelligence test performance of lower-class black children. *Journal of Consulting and Clinical Psychology, 46*, 1538-1539.

96. Terrell, F., Terrell, S. L., & Taylor, J. (1980). Effects of race of examiner and type of reinforcement on the intelligence test performance of lower-class black children. *Psychology in the Schools, 17*, 270-272.

97. Terrell, F., Terrell, S. L., & Taylor, J. (1981). Effects of type reinforcement on the intelligence test performance of retarded black children. *Psychology in the Schools, 18*, 225-227.

98. Tindal, G., Almond, P., Heath, B., & Tedesco, M. (1998). *Single subject research using audio cassette read aloud in math.* Manuscript submitted for publication, University of Oregon.

99. Tindal, G., Glasgow, A., Helwig, B., Hollenbeck, K., & Heath, B. (1998). *Accommodations in large scale tests for students with disabilities: An investigation of reading math tests using video technology.* Unpublished manuscript with Council of Chief State School Officers, Washington, DC.

100. Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children, 64* (4), 439-450.

101. Tindal, G., Hollenbeck, K., Heath, B., & Almond, P. (1998). *The effect of using computers as an accommodation in a statewide writing test.* Manuscript submitted for publication, University of Oregon.

102. Trimble, S. (1998). *Performance trends and use of accommodations on a statewide assessment* (Maryland/Kentucky State Assessment Series Rep. No. 3). Minneapolis, MN: National Center on Educational Outcomes.

103. Vacc, N. N. (1987). Word processor versus handwriting: A comparative study of writing samples produced by mildly mentally handicapped students. *Exceptional Children, 54*, 156-165.

104. Varnhagen, S., & Gerber, M. M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly, 7*, 266-270.

105. Veit, D. T., & Scruggs, T. E. (1986). Can learning disabled students effectively use separate answer sheets? *Perceptual and Motor Skills, 63*, 155-160.

106. Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53-79.

107. Watkins, M. W., & Kush, J. C. (1988). Assessment of academic skills of learning disabled students with classroom microcomputers. *School Psychology Review, 17*, 81-88.

108. Weaver, S. M. (1993). *The validity of the use of extended and untimed testing for postsecondary students with learning disabilities (extended testing).* Unpublished doctoral dissertation, University of Toronto, Toronto.

109. Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment, 1*, 131-152.

110. Westin, Tim (April 1999). *The validity of oral presentation in testing.* Montreal, CANADA: American Educational Research Association.

111. Wheeler, L. J., & McNutt, G. (1983). The effects of syntax on low-achieving students' abilities to solve mathematical word problems. *The Journal of Special Education, 17* (3), 309-315.

112. Whinnery, K. W., & Fuchs, L. S. (1993). Effects of goal and test-taking strategies on the computation performance of students with learning disabilities. *Learning Disabilities Research & Practice, 8*, 204-214.

113. Willis, J., & Shibata, B. (1978). A comparison of tangible reinforcement and feedback effects on the WPPSI I. Q. scores of nursery school children. *Education and Treatment of Children, 1*, 31-45.

114. Young, R. M., Bradley-Johnson, S., & Johnson, C. M. (1982). Immediate and delayed reinforcement on WISC-R performance for mentally retarded students. *Applied Research in Mental Retardation, 3*, 13-20.

115. Ziomek, R. L., & Andrews, K. M. (1996). *Predicting the college grade point averages of special-tested students from their ACT assessment scores and high school grades.* Iowa City, IA: ACT.

References in Support of the Research on Test Changes

Anderson, N. E., Jenkins, F. F., & Miller, K. E. (1995). *NAEP inclusion criteria and testing accommodations*. Princeton, NJ: Educational Testing Service Unpublished manuscript.

Bond, L.A., & Roeber, E.D. (1995). *The status of state student assessment programs in the United States*. Oak Brook, IL: North Central Regional Educational Laboratory (NCREL) and the Council of Chief State School Officers (CCSSO).

Brigance, A.H. (1978). *Inventory of Early Development.* Woburn, MA: Curriculum Associates.

Brown, J. I., Fishco, V. V., & Hanna, G. S. (1981-1993). *Nelson-Denny Reading Test*, Forms G and H. Riverside, CA: Riverside Publishing CO.

Chin-Chance, S. A., Gronna, S. S., & Jenkins, A. A. (1996). *Assessing special education students in a norm-referenced statewide testing program: Hawaii State Department of Education*. Unpublished manuscript for the Council of Chief State School Officers.

Chiu, C. & David Person , D. (1998). *Bibliography of Empirical Studies on Test Accommodations for Special Education Students*. Michigan State University.

Council of Chief State School Officers. (1996). *The status of state student assessment programs in the United States*. Washington, DC: Author.

Deno, S.L., & Mirkin, P.K. (1980). Data-based IEP develop: An approach to substantive compliance. *Teaching Exceptional Children, 12* (3), 92-97.

Erickson, R.N., & Thurlow, M.L.(1996). *State special education outcomes 1995*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Erickson, R.N., Thurlow, M.L., & Thor, K. (1995). *1994 state special education outcomes*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Erickson, R.N., Thurlow, M.L., & Ysseldyke, J.E. (1996). *Fractured fractions: Determining the participation rates for students with disabilities in statewide assessment programs*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Evans, F. R. (1980). *A study of relationships among speed and power aptitude test scores , and ethnic identity.* (College Board Report RDR 80-81, No. 2, and ETS RR 80-22). Princeton, NJ: Educational Testing Service.

Evans, F. R., and Reilly, R. R. (1972a). *The LSAT Speededness Study – Revisited.* Law School Admission Test Council Annual Report. Princeton, NJ: Educational Testing Service.

Evans, F. R., and Reilly, R. R. (1972b). A study of speededness as a source of test bias. *Journal of Educational Measurement, 9*, 123-131.

Evans, F. R., and Reilly, R. R. (1973). A study of test speededness as a potential source of bias in quantitative score of the Admission Test for Graduate Study in Business. *Research in Higher Education, 1*, 173-183.

Fredericksen, J. R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.

Fuchs, L.S., Fuchs, D., & Hamlett, C.L. (1990). Curriculum-based measurement: A standardized, long-term goal approach to monitoring student progress. *Academic Therapy, 25*, 615-632.

Fuchs, L.S., Fuchs, D., Hamlett, C.L., & Stecker, P.M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28*, 617-641.

Gajria, M., Salend, S. J., & Hemrick, M. A. (1994). Teacher acceptability of testing modifications for mainstreamed students. *Learning Disabilities: Research & Practice, 9*(4), 236-243.

Gulliksen, H. O. (1950). *Theory of mental tests.* New York: John Wiley & Sons.

Hartman, R. C., & Redden, M. R. (1985). *Measuring student progress in the classroom: A guide to testing and evaluating progress of students with disabilities (1985-86 ed.).* Washington, DC: Department of Education (ERIC Document Reproduction Service No. 295 403).

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *The Journal of Special Education, 32*(3),175-183.

Jayanthi, M., Epstein, M. H., Polloway, E. A., & Brusuck, W. D. (1996). A national survey of general education teachers' perceptions of testing adaptations. *The Journal of Special Education, 30*(1), 99-115.

Kavale, K.A., & Reece, J.H. (1992). The character of learning disabilities. *Learning Disability Quarterly, 15*, 74-94.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky (CSE Technical Report No. 431).* Los Angeles, CA: Center for Research on Standards and Student Testing.

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11-16.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*(2), 179-189.

Linn, R. L. (1993). Educational assessment expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-6.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-23.

McDonnell, L.M., McLaughlin, M.W., & Morison, P. (1997). *Educating one and all: Students with disabilities and standards-based reform.* Washington, DC: National Academy Press.

McGrew, K.S., Thurlow, M.L., Shriner, J.G., & Spiegel, A.N. (1992). *Students with disabilities in national and state data collection programs.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational Measurement-Third Edition* (pp. 13-104). New York: Macmillan.

Messick, S. (1995a). Special Issue: Values and standards in performance assessment: Issues, findings and viewpoints. *Educational Measurement: Issues and Practice, 14*(4), 4.

Messick, S. (1995b). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4),

Miller, M. D. (1996). *Generalizability in Connecticut.* Unpublished manuscript written for Council of Chief State School Officers.

Miller, M. D. (1996). *Generalizability in Connecticut.* Unpublished manuscript written for Council of Chief State School Officers.

Miller, M. D., Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25*(3), 205-219.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4), 287-304.

Muers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17*(3), 347-352.

National Council of Teachers of Mathematics (1989). *Standards.* Washington, D.C.

National Council on Education Standards and Testing (1992). *Raising standards for American education.* Washington, DC: US Government Printing Office.

National Council on Measurement in Education (1995). *Code of professional responsibilities in educational measurement.* Authors.

NCEO (1996). State project highlights – Increasing the participation of students with disabilities. *Datalinks*, August.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice, 11*(1), 9-15.

Nolet, V., Tindal, G. (1996). Serving students in middle school content classes: a heuristic study of critical variables linking instruction and assessment. *The Journal of Special Education, 29*(4), 414-432.

O'Sullivan & Chalnick, (1991). Measurement-related coursework requirements for teacher certification and recertification. *Educational Measurement: Issues and Practice, 10*(1), 17-19, 23.

Odden, A. (1990). Educational indicator in the United States: The need for analysis. *Educational Researcher, 19*(4), 24-29.

Phillips, S. E. (1994). High-stakes testing accommodations: validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93-120.

Phillips, S.E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93-120.

Porter, A. (1995). The uses and misuses of opportunity to learn standards. *Educational Researcher*, 21-27.

Porter, A. (1993). School delivery standards. *Educational Researcher, 22*(5), 24-30.

Poteet, J. A., Choate, J. S., & Stewart, S. C. (1993). Performance assessment and special education: Practices and Prospects. *Focus on Exceptional Children, 26*(1), 1-16.

Potter, P., & Mirkin, P. K. (1982). *Instructional planning and implementation practices of elementary and secondary resource room teachers*. [Research Report no. 65]. Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Reckase, M. D. (1995). Portfolio assessment: a theoretical estimate of score reliability. *Educational Measurement: Issues and Practice, 14*(1), 12-14.

Salvia, J. & Ysseldyke, J. (1998). *Assessment (7th edition)*. Boston: Houghton Mifflin.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*(1), 3-6, 12.

Ruiz-Primo, Baxter, & Shavelson (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41-54.

Shepard, L. (1983). The role of measurement in education policy: Lessons from the identification of learning disabilities. *Educational Measurement: Issues and Practice, 2*(3), 4-8.

Shriner, J. G., Gilman, C. J., Thurlow, M. L., Ysseldyke, J. E. (1994-95). Trends in state assessment of educational outcomes. *Diagnostique, 20*(1-4), 101-119.

Shriner, J.G., & Thurlow, M.L. (1992). *State special education outcomes 1991*. Minneapolis, MN: University of Minnesota National Center on Educational Outcomes.

Siskind, T. G. (1993a). Modifications in statewide criterion-referenced testing programs to accommodate pupils with disabilities. *Diagnostique, 18*(3), 233-249.

Siskind, T. G. (1993b). Teachers' knowledge about test modifications for students with disabilities. *Diagnostique, 18*(2), 145-157.

Smith, S.W. (1990). Comparison of individualized education programs (IEPs) of students with behavioral disorders and learning disabilities. *The Journal of Special Education, 24*, 85-99.

Smith, S.W., & Simpson, R.L. (1989). An analysis of individualized education programs (IEPs) for students with behavior disorders. *Behavioral Disorders, 14*, 107-116.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10*(1), 7-12.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with learning disabilities: A review of the literature*. Mpls., MN: University of Minnesota National Center on Educational Outcomes.

Thurlow, M., Erickson, R., Spicuzza, R., Vieburg, K., & Ruhland, A. (1996). *Accommodations for students with disabilities: Guidelines from states with graduation exams (Minnesota Report No. 5)*. Mpls., MN: University of Minnesota National Center on Educational Outcomes.

Thurlow, M., Hurley, C., Spicuzza, R., & Sawaf, H. E. (1996). *A review of the literature on testing accommodations for students with disabilities*. State Assesment Series – Minnesota Report 9. University of Minnesota National Center on Educational Outcomes.

Thurlow, M.L., Scott, D.L., & Ysseldyke, I.E. (1995a). *A compilation of states' guidelines for accommodations in assessments for students with disabilities* (Synthesis Report 18). Minneapolis, MN: University of Minnesota National Center on Educational Outcomes.

Thurlow, M.L., Scott, D.L., & Ysseldyke, J.E. (1995b). *A compilation of states' guidelines for including students with disabilities in assessments* (Synthesis Report 17). Minneapolis, MN: University of Minnesota National Center on Educational Outcomes.

Thurlow, M.L., Ysseldyke, J.E., & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education, 16*(5), 260-270.

Tindal, G. (1997). Performance assessment. In R. Taylor (Ed.), *Assessment of individuals with mental retardation*. San Diego, CA: Singular Publishing Group.

Tindal, G. (1998a). Assessment in learning disabilities with a focus on curriculum-based measurement (pp. 35-66). In J. Torgeson & B. Wong (Eds.). *Learning about Learning Disabilities*. San Diego, CA: Academic Press.

Tindal, G. (1998b). Issues in performance assessment for students with disabilities: Inclusion, technical adequacy, and interpretation of performance outcomes (pp. 73-102). In R. Taylor (Ed.) *Assessment of individuals with mental retardation*. San Diego: Singular Press.

Tindal, G. (l998c). *Models for understanding task comparability in accommodated testing*. Eugene, OR: Behavioral Research and Teaching.

Tindal, G., & Nolet, V. (1996). Serving students in middle school content classes: A heuristic study of critical variables linking instruction and assessment. *The Journal of Special Education*, *29*(4), 414-432.

Tindal, G., Fuchs, L.S., Fuchs, D., Shinn, M., Deno, S.L., & Germann, G. (l985). Empirical validation of criterion-referenced test. *Journal of Educational Research, 78*, 203-209.

Wesson, C.L. (l991). Curriculum-based measurement and two models of follow-up consultation. *Exceptional Children, 57*, 246-257.

Wesson, C. L., Deno, S. l., & Mirkin, P. K. (1982). *Research on developing and monitoring progress on IEP goals: A review of the literature.* [Monograph No. 8]. Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Willingham, W.W. (l989). Standard testing conditions and standard score meaning for handicapped examinees. *Applied Measurement in Education, 2*(2), 97-103.

Winter, P.C. (1996). *State research on performance assessment.* Washington DC: Council of Chief State School Officers.

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practices, 8*(3), 5-10.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education*, *42*(1), 37-42.

Ysseldyke, J., Thurlow, M., McGrew, K., & Shriner, J. (1994). *Recommendations for making decisions about the participation of students with disabilities in statewide assessment programs*, [Synthesis Report No. 15]. Minneapolis: University of Minnesota National Center on Educational Outcomes.

Ysseldyke, J., Thurlow, M., McGrew, K., & Vanderwood, M. (1994). *Making decisions about the inclusion of students with disabilities in large-scale assessments* (Synthesis Report 13). Minneapolis: University of Minnesota National Center on Educational Outcomes.

Ysseldyke, J.E., Erickson, R., Gabrys, R., Haigh, J., Trimble, S., & Gong, B. (1996). *A comparison of state assessment systems in Maryland and Kentucky with a focus on participation of students with disabilities*. Unpublished manuscript.

Ysseldyke, J.E., Thurlow, M.L., McGrew, K.S., & Shriner, J.G. (1994). *Recommendations for making decisions about participation of students with disabilities in statewide assessment programs* (Synthesis Report 15). Minneapolis: University of Minnesota National Center on Educational Outcomes.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☑ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").